A Fluid Analysis of a Utility-Based Wireless Scheduling Policy

Peijuan Liu, *Member, IEEE*, Randall A. Berry, *Member, IEEE*, and Michael L. Honig, *Fellow, IEEE*

***Abstract*—In this paper, we consider packet scheduling for the downlink in a wireless network, where each packet's service preferences are captured by a utility function that depends on the total delay incurred. The goal is to schedule packet transmissions to maximize the total utility. In this setting, we examine a simple gradient-based scheduling algorithm called the $\dot{U}R$-rule, which is a type of generalized $c\mu$-rule $(Gc\mu)$ that takes into account both a user's channel condition and derived utility when making scheduling decisions. We study the performance of this scheduling rule for a draining problem, where there is a given set of initial packets and no further arrivals. We formulate a "large system" fluid model for this draining problem where the number of packets becomes large while the packet-size decreases to zero, and give a complete characterization of the behavior of the $\dot{U}R$ scheduling rule in this limiting regime. Comparison with simulation results show that the fluid limit accurately predicts the corresponding behavior of finite systems of interest. We then give an optimal control formulation for finding the optimal scheduling policy for the fluid draining model. Using Pontryagin's minimum principle, we show that, when the user rates are chosen from a TDM-type of capacity region, the $\dot{U}R$ rule is in fact optimal in many cases. Sufficient conditions for optimality are also given. Finally, we consider a general capacity region and show that the $\dot{U}R$ rule is optimal only in special cases.**

***Index Terms*—Fluid model, optimal control, packet scheduling, utility function, wireless scheduling.**

I. Introduction

Efficient scheduling algorithms are recognized as a key component for providing high speed wireless data services. A basic characteristic of wireless systems is that channel quality will vary across the user population, enabling different users to receive data at different rates. There has been much interest in "channel-aware" scheduling algorithms that exploit these variations in channel quality to improve system performance (e.g., [1]–[13], [15]–[17]). An important consideration for such scheduling approaches is balancing the overall system performance with each user's quality of service (QoS) requirements. For example, in a time-division multiplexing (TDM) system that transmits to one user at a time, the overall throughput is maximized by always transmitting to the user with the best channel. However, this approach can result in poor performance for users with poor channel quality. This problem is especially prominent in a low-tier mobility environment where channel conditions vary slowly with time. To address these considerations, various "fair" scheduling approaches have been considered, such as the *proportional fair* algorithm proposed for the CDMA 1xEV-DO system [18], [19]. Other approaches for addressing fairness include emulating the generalized processor sharing (GPS) model [3] or imposing various "resource-sharing" constraints on the system [12].

In this paper, we consider a utility-based scheduling framework, where each packet has a utility function (which can vary across packets) that indicates the benefit from receiving the packet after a certain delay. The scheduling policy then attempts to maximize the total system utility; in this way, the utility functions can be used to balance fairness and efficiency. We consider a simple gradient-based scheduling policy, which we call the $\dot{U}R$ scheduling rule [16], [17]. Here $\dot{U}$ represents the marginal utility associated with scheduling the packet, and R is the achievable rate, which is related to the channel quality.[1] This policy makes decisions based only on the instantaneous values of these parameters, and so requires no knowledge about the fading statistics or user traffic.

We consider scheduling for the downlink of a single cell in an environment where the channel gain to each user is known and fixed over the time-scale of interest.[2] This assumption is reasonable in a slow fading environment and may be appropriate, for example, for fixed wireless access or a broadband satellite system. Note that in this setting, issues of "opportunistic" scheduling do not arise, e.g., [12], [13], [21]. One reason we focus on this time-invariant model is that it highlights the possible disparity among users when certain users' channel conditions are consistently inferior to others. We have also shown in [16] that the performance benefits of the $\dot{U}R$ scheduling policy are the most prominent in an environment with static channel gains. The basic model considered here also applies to scheduling in other multi-class queueing systems where different classes have different service rates, for example, in a wire-line network where different classes have different packet lengths.

Manuscript received June 30, 2004. This work was supported by the Northwestern-Motorola Center for Communications, and by the NSF under Grants CCR-9903055 and CAREER award CCR-0238382. This paper was presented in part at the 40th Annual Allerton Conference on Communications, Control, and Computing, Monticello, IL, and the 2004 IEEE Conference on Decision and Control, Paradise Island, the Bahamas. The work of P. Liu was performed while she was with the Department of Electrical Engineering and Computer Science, Northwestern University.

P. Liu is with Motorola Labs, Schaumburg, IL 60196 USA (e-mail: Peijuan.Liu@motorola.com).

R. A. Berry and M. L. Honig are with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208 USA (e-mail: rberry@ece.northwestern.edu; mh@ece.northwestern.edu).

Communicated by G. Sasaki, Associate Editor Communication Networks.

Digital Object Identifier 10.1109/TIT.2006.876246

[1] The marginal utility can be interpreted as a "bid" price and reflects the urgency of transmitting the packet when $U(\cdot)$ is a function of delay.

[2] We note however, that the $\dot{U}R$ scheduling policy can be easily applied to a system with time-varying channels [16].

0018-9448/$20.00 © 2006 IEEE

We analyze the performance of the $\dot{U}R$ policy for a draining model, where there is an initial set of packets to send, each with an initial delay, and no new arrivals occur. We formulate a fluid limit for this problem, where the number of initial packets increases, while the packet size decreases to zero. A complete characterization of the performance of the $\dot{U}R$ scheduler is given for the fluid system. We then consider the optimal scheduling policy for a fluid system with two classes of users; this can be formulated as a continuous-time optimal control problem. Using Pontryagin's minimum principle, we show that in certain cases the $\dot{U}R$ scheduler is optimal, i.e., it maximizes the total utility. We also show that the optimality of the $\dot{U}R$ rule depends in part on the underlying physical layer capacity region. For a TDM type of capacity region, the $\dot{U}R$ rule is optimal for a broad class of utility functions; for a general capacity region, the $\dot{U}R$ rule is optimal only in some special cases.

The $\dot{U}R$ policy is equivalent to the *generalized* $c\mu(Gc\mu)$ rule introduced by Van Meighem in [22] for a single-server multi-class queueing system with general convex delay costs.[3] In [22] it is shown that the $Gc\mu$ rule is asymptotically optimal in the heavy traffic regime. The heavy traffic optimality of a $Gc\mu$ rule for a system with multiple flexible servers is shown in [23] under the assumption of "complete resource sharing." Here we do not consider the heavy traffic regime, but instead analyze the performance and optimality of this rule for the fluid draining problem previously discussed. A different fluid "rush hour" model has been studied in [24]; the authors argue that a $Gc\mu$ rule is often optimal in this setting as well. Optimal control of fluid models for other queueing systems (typically with linear costs) has also received some attention, e.g., [26].

We allow the utility to be an arbitrary concave decreasing function of delay. In the special case of linear utilities, the $\dot{U}R$ rule reduces to the well-known $c\mu$-rule which is known to be optimal in a variety of settings (e.g., [27]–[29]). With quadratic utilities, the $\dot{U}R$ rule is equivalent to the "MaxWeight" policies studied in [1], [5], [25]. The "MaxWeight" scheduling rules are stabilizing policies in a variety of settings, e.g., [1], [5] and also exhibit several optimal properties in the heavy traffic regime [25]. Several other fair scheduling approaches, such as the proportional fair rule, can be viewed in terms of utilities that depend on each user's throughput averaged over a sufficiently long period. In that setting, algorithms similar to the $\dot{U}R$ rule can be used to maximize the total utility [7], [14].

The remainder of the paper is organized as follows. In Section II, we describe the system model and motivate the $\dot{U}R$ rule. In Section III, we analyze the performance for a system with K classes of packets, where each class is differentiated by its utility function and achievable transmission rate. We formulate a fluid limit, and characterize the associated performance. In Section IV, we extend our analysis to a limiting system with an infinite number of classes, i.e., we allow an arbitrary distribution for rates across packets. In Section V, we present an optimal control formulation for finding the optimal scheduling policy given a TDM capacity region. For a broad class of utility functions, it is shown that the $\dot{U}R$ rule is optimal. In Section VI, we consider the optimal scheduling policy for a general capacity region and give necessary conditions for the $\dot{U}R$ rule to be optimal. We observe that the $\dot{U}R$ rule satisfies those conditions only in special cases.

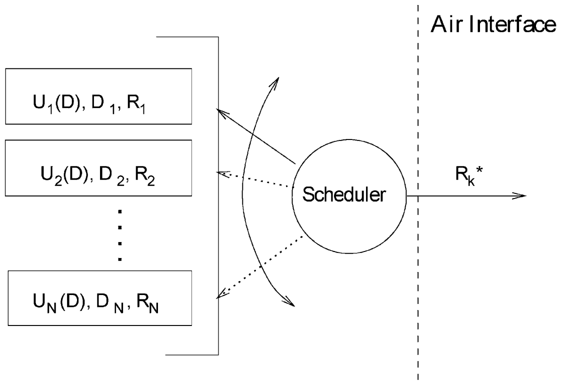


Fig. 1. System model with N packets waiting to be scheduled.

II. Utility-Based Downlink Scheduling

We consider a basic model for downlink scheduling from a single transmitter, such as a base station in a cellular network or an access point in a wireless LAN. We initially consider a TDM system where the transmitter sends to one user at a time, as in the CDMA 1xEV-DO standard [18], [20]. We also discuss the case where multiple users may be scheduled simultaneously and assigned rates determined by a given physical layer "capacity region." This can model systems such as CDMA 1xEV-DV[20] where a subset of users are scheduled in each time-slot, and the available spreading codes and transmission power are then allocated among the scheduled users to determine their transmission rates. This could also model the case where the downlink is modeled as a Gaussian broadcast channel and any set of achievable rates in the broadcast capacity region can be used.

Consider a scheduling instant when N packets are queued at the base station waiting for transmission. In a TDM system, each packet i is associated with a transmission rate R_i that reflects the corresponding channel quality to the intended receiver. The scheduler decides which packet to transmit based on the transmission rate, along with the packet's utility function, and the current delay (see Fig. 1). The utility received by sending the ith packet, $U_i(D_i)$, is a decreasing, concave function of its total delay, D_i (i.e., the packet's sojourn time). Let $W_i(t)$ denote the waiting time of the ith packet at time t. During each scheduling interval, if the scheduler decides to transmit the kth packet, then that packet is sent over the air interface at rate R_k. For simplicity, each packet is assumed to contain L bits including any overhead. The goal is then to schedule the packets to maximize the average utility rate

$$U_{\text{avg}} = \lim_{T\to\infty} \frac{1}{T} \sum_{i=1}^{N(T)} U_i(D_i) \tag{1}$$

[3] A utility U that is a function of delay is equivalent to a delay cost of $-U$.

where $N(T)$ denotes the total number of packets served up to time T, and D_i represents the total delay experienced by packet i.

We consider a simple gradient-based scheduling policy. This policy attempts to schedule a packet from the class that results in the largest first-order change in the total utility rate. In a TDM system, if the scheduler transmits to the ith packet, followed by the jth packet, the change in total utility is given by

$$\Delta U_{i,j} := U_i\left(W_i(t) + \frac{L}{R_i}\right) + U_j\left(W_j(t) + \frac{L}{R_i} + \frac{L}{R_j}\right).$$

Approximating $U_i(\cdot)$ by a first-order Taylor series around $W_i(t)$ we have

$$\begin{aligned}\Delta U_{i,j} &\approx \tilde{\Delta} U_{i,j}\\ &:= U_i(W_i(t)) + \dot{U}_i(W_i(t))\frac{L}{R_i}\\ &\quad + U_j(W_j(t)) + \dot{U}_j(W_j(t))\left(\frac{L}{R_i} + \frac{L}{R_j}\right).\end{aligned}$$

Likewise, transmitting in the reverse order yields

$$\begin{aligned}\Delta U_{j,i} &\approx \tilde{\Delta} U_{j,i}\\ &:= U_i(W_i(t)) + \dot{U}_i(W_i(t))\left(\frac{L}{R_i} + \frac{L}{R_j}\right)\\ &\quad + U_j(W_j(t)) + \dot{U}_j(W_j(t))\frac{L}{R_j}.\end{aligned}$$

Simplifying the preceding expressions gives the following scheduling rule, which favors packet i over packet j if $\tilde{\Delta} U_{i,j} > \tilde{\Delta} U_{j,i}$ for $j \neq i$.

$\dot{U}R$ Scheduling Rule: Schedule user i^* such that

$$i^* = \arg\max_i |\dot{U}_i(W_i(t))| R_i \tag{2}$$

where ties are broken arbitrarily.

Here we have used that since $U_i(\cdot)$ is decreasing, $\dot{U}_i(W_i(t))$ is negative.

In the general setting where multiple transmissions are allowed, let $\mathbf{r} = \{r_1, \ldots, r_N\}$ be the transmission rate vector for all packets. A natural generalization of (2) is for the scheduler to choose a rate vector, $\mathbf{r}$, such that

$$\mathbf{r} = \arg\max_{\mathbf{r}\in\mathcal{C}} \sum_{i=1}^{N} |\dot{U}_i(W_i)| r_i \tag{3}$$

where ties are broken arbitrarily. The set $\mathcal{C}$ denotes the N-dimensional capacity region of feasible rates. For a TDM scheme

$$\mathcal{C}_{\text{TDM}} := \left\{\begin{array}{c} \{\ R_1,\ 0,\ 0,\ \ldots,\ 0\ \}\\ \{\ 0,\ R_2,\ 0,\ \ldots,\ 0\ \}\\ \ldots,\\ \{\ 0,\ \ldots,\ 0,\ 0,\ R_N\ \}\end{array}\right\}. \tag{4}$$

In this case, the rule specified by (3) reduces to (2).

III. K-Class System

A. System Model

We consider a draining problem where a group of packets are present at time $t = 0$ and no new arrivals occur. Each packet has a random initial delay. This could model a system with batch arrivals, where the time between arrivals is sufficiently long to drain the previous batch. Each packet is associated with a randomly chosen transmission rate and a utility function. The goal is to drain these packets while maximizing the average utility per packet.

We first consider a TDM system with K classes of packets; each class corresponds to packets with the same feasible transmission rate and service requirements.[4] Specifically, for $i = 1, \ldots, K$, the base station can transmit class i packets with transmission rate R_i. We assume that $R_1 \geq R_2 \geq \cdots \geq R_K$ and that these rates are fixed over the time horizon of interest.

Initially assume there are N packets in the system and no new arrivals occur. Each packet is independently assigned to class i with probability $p_i (i = 1, \ldots, K)$. Let N_i denote the number of class i packets; this is a random variable with expected value $[N_i] = p_i N$. The system is to be emptied by transmitting all of the N packets. The time required to drain the system with any work-conserving (nonidling) scheduling rule is given by

$$T_f = \sum_{i=1}^{K} \frac{N_i L}{R_i}. \tag{5}$$

This is independent of the order in which packets are served. However, the service order does influence the delay incurred by the individual packets, and hence the derived utility.

We assume that each packet has an initial delay at time $t = 0$. This reflects the delay experienced by the packets prior to time $t = 0$ and could include, for example, the delay incurred in forwarding the packet to the base station. For $k = 1, \ldots, N_i$, we denote the initial delay of the kth packet of class i by $W_{i,k}(0)$. If this packet is transmitted after t seconds, then the total delay incurred is

$$D_{i,k} = W_{i,k}(0) + t + \frac{L}{R_i},$$

where L/R_i is the transmission time.

The utility associated with each class i packet served is given by $U_i(D_{i,k})$. The utility per packet generated by a given schedule is

$$U_{\text{avg}} = \frac{1}{N}\sum_{i=1}^{K}\sum_{k=1}^{N_i} U_i(D_{i,k}).$$

Notice that this depends on the initial delays for the packets in each class.

For a given initial delay distribution, a schedule of packet transmissions is defined to be *optimal* if it maximizes U_{avg}. Consider the special case where $U_i(x) = -x$ for $i = 1, \ldots, K$, and thus maximizing U_{avg} becomes equivalent to minimizing the average delay per packet. In this case, the optimal schedule is to transmit packets in decreasing order of transmission

[4] For the problem considered here all of the packets in a given class can be directed to one user or several users with similar channels/requirements.

rates; within each class, the order in which packets are transmitted does not effect U_{avg}. This can be shown using a simple interchange argument. Next suppose that the utility $U_i(\cdot)$ is strictly concave for each i. Then it can be shown that the optimal scheduler transmits packets within each class in longest-delay-first order, i.e., if $W_{i,k}(0) > W_{i,\tilde{k}}(0)$, then packet k is transmitted before packet $\tilde{k}$. Therefore, in the following we will only consider scheduling among the head-of-line packet within each class. Even with this characterization, there are $\prod_{k=1}^{K} \binom{N - \sum_{i=0}^{k-1} N_i}{N_k}$ possible schedules from which to choose. In the case of linear utilities, the $\dot{U}R$ rule becomes a $c\mu$ rule; hence, we have the following.

Proposition 1: If $U_i(x) = -\beta_i x$ for $i = 1, \ldots, K$ and $\beta_i > 0$, then the $\dot{U}R$ scheduling rule maximizes the utility per packet.

The proof of this result with zero initial delays is given in [28]. It is easy to show that adding initial delays does not affect the scheduling decisions, and that the $\dot{U}R$ rule is still optimal.

B. Fluid Limit

To analyze the performance of scheduling policies for the draining problem, we consider a type of fluid limit for the system. In this section, we describe this limit for an arbitrary scheduling rule. In the next section, we consider the limiting behavior of the $\dot{U}R$ scheduling rule.

We scale up the number of packets and decrease the packet size, while keeping a fixed load (in bits).[5] Formally, we consider a sequence of systems indexed by $N = 1, 2, \ldots$; in the Nth system there are initially N packets in total with packet length L normalized so that $NL = 1$.[6] With this scaling, T_f in (5) will converge to $\sum_{i=1}^{K} p_i/R_i$ almost surely, by the strong law of large numbers. As noted previously, the performance of a scheduler depends on the initial packet delays. For each class i, we assume that $\{x_{i,k}\}_{k=1}^{\infty}$ is also sequence of *i.i.d.* random variables, with cumulative distribution function (*cdf*) $G_i(w) = \Pr(x_{i,k} \le w)$ and probability density function (*pdf*) $g_i(w)$. Let $W_{i,k}(0) = x_{i,k}$ for $k = 1, \ldots, N_i$, i.e., the initial delays for class i packets in the Nth system are set to be the first N_i components of this sequence. For simplicity, we further assume that $g_i(w) > 0$ if and only if $w \in [D_i^{\min}, D_i^{\max}]$, where $D_i^{\min} \ge 0$ and $D_i^{\max} < \infty$ are lower and upper bounds on the initial delay, respectively.[7]

Let $\mathcal{N}_i^N(t)$ denote the number of class i packets remaining at time t in the Nth system (for a given scheduling policy), and let

$$f_i^N(t) = \frac{\mathcal{N}_i^N(t)}{N_i}$$

be the fraction of class i packets remaining at time t. Likewise, let $\tau_i^N(t)$ denote the amount of time in $[0, t)$ during which the transmitter serves packets from class i. Between times t and $t + \delta t$, the change in $f_i^N(t)$ can be bounded as

$$\begin{aligned} &\frac{-\left[\tau_i^N(t+\delta t) - \tau_i^N(t)\right]\frac{R_i}{L} - 1}{N_i \delta t} \\ &\quad \le \frac{f_i^N(t+\delta t) - f_i^N(t)}{\delta t} \\ &\quad \le \frac{-\left[\tau_i^N(t+\delta t) - \tau_i^N(t)\right]\frac{R_i}{L} + 1}{N_i \delta t}. \end{aligned} \tag{6}$$

For a finite N, the preceding quantities depend on the initial delays and the number of packets in each class, and hence are random. For the scheduling policies of interest, we assume that as $N \to \infty, \tau_i^N(t)$ converges almost surely to a deterministic limit $\tau_i(t)$.

As $N \to \infty, L = 1/N \to 0$ and $N_i/N \to p_i$. Therefore, from (6) it follows that $f_i(t) = \lim_{N\to\infty} f_i^N(t)$ exists and satisfies

$$\frac{f_i(t+\delta t) - f_i(t)}{\delta t} = \frac{-[\tau_i(t+\delta t) - \tau_i(t)]R_i}{p_i \delta t}.$$

Next, letting $\delta t \to 0$, we have[8]

$$\dot{f}_i(t) = -\frac{\alpha_i(t) R_i}{p_i} \tag{7}$$

where $\alpha_i(t) := \dot{\tau}_i(t)$. Notice that both $f_i(t)$ and $\tau_i(t)$ are monotonic functions of t and hence the preceding derivatives exist except possibly on a set of measure zero. At those values of t where $\tau_i(t)$ is not differentiable, we set $\alpha_i(t)$ to be the right derivative. Therefore $\alpha_i(t)$ is right continuous [31].

In the limit, the base station can transmit arbitrarily many packets in any time interval $[t, t + \delta t)$, but only a finite fraction of the initial packets. The fraction of unserved packets in class i at time $t \in [0, T_f)$ is then given by

$$f_i(t) = 1 - \int_0^t \frac{\alpha_i(\tau) R_i}{p_i} d\tau. \tag{8}$$

The quantity $\alpha_i(t)$ can be interpreted as the fraction of the base station's resources devoted to class i packets at time t. If $\alpha_i(t) = 1$, then only class i packets are served. In general, $\alpha_i(t)$ can take on any value in $[0, 1]$ and must satisfy $\sum_{i=1}^{K} \alpha_i(t) \le 1$ for each time t. For a nonidling policy, $\sum_{i=1}^{K} \alpha_i(t) = 1$ for all $t \in [0, T_f)$. At each time t, the scheduling algorithm specifies $\alpha_i(t)$. Equivalently, for this limiting TDM system, we can view the scheduler as selecting rates $r_i(t) = \alpha_i(t) R_i$ from the capacity region given by $\mathcal{C}_{TS} = \{\mathbf{r} : \sum_{i=1}^{K} \frac{r_i}{R_i} = 1\}$. This is a K-dimensional simplex with corner points given by the set $\mathcal{C}_{\text{TDM}}$ defined in (4). This interpretation generalizes directly to other capacity regions $\mathcal{C}$ as in (3). In this case, the scheduler selects rates $\mathbf{r}(t) \in \mathcal{C}$ and the fraction of unserved packets within each class evolves according to $\dot{f}_i(t) = -\frac{r_i(t)}{p_i}$.

[5] This can also be viewed in terms of the "usual" fluid scaling for queuing systems, where time and space are linearly scaled by N, and where as N increases, the number of initial packets also increases.

[6] There is no loss in generality in assuming that the product NL is normalized to 1. A system with $NL = c$ with $c \ne 1$ and rates $\{R_i\}_{i=1}^{K}$, can be easily shown to behave equivalently to a system with $NL = 1$ and rates $\{R_i/c\}_{i=1}^{K}$.

[7] This can be extended to $M < \infty$ disjoint compact support ranges $[D_i^{\min_1}, D_i^{\max_1}] \cup [D_i^{\min_2}, D_i^{\max_2}] \cup \cdots \cup [D_i^{\min_M}, D_i^{\max_M}]$. In that case, treat this class as M classes with the same transmission rate. The initial delay distribution for each sub-class m then has support $[D_i^{\min_m}, D_i^{\max_m}], m = 1, \ldots, M$.

[8] We use the notation $\dot{f}(x)$ to indicate the first derivative of f with respect to its argument, i.e., $\frac{df}{dx}$.

As an example of the preceding scaling, consider a TDM system with two equally loaded classes ($p_1 = p_2 = 1/2$) using a round robin scheduling policy that alternates between scheduling class 1 and class 2 packets. In this case, for the Nth system we have

$$\left(\frac{t}{\frac{L}{R_1}+\frac{L}{R_2}} - 1\right)\frac{L}{R_1} \le \tau_1^N(t) \le \left(\frac{t}{\frac{L}{R_1}+\frac{L}{R_2}} + 1\right)\frac{L}{R_1}.$$

Hence, as $N \to \infty, \tau_1^N(t)$ converges to $\tau_1(t) = \frac{R_2 t}{R_1+R_2}$, so that $\alpha_1(t) = \frac{R_2}{R_1+R_2}$ and $\alpha_2(t) = \frac{R_1}{R_1+R_2}$.

Next, we turn to the packet delays in the limiting system. For a given realization of $\{W_{i,k}(0)\}_{k=1}^{N_i}$, let $G_i^N(w)$ denote the empirical distribution of the initial delays for type i packets in the Nth system, i.e.,

$$G_i^N(w) = \frac{|\{k \le N_i : W_{i,k}(0) \le w\}|}{N_i}$$

where $|\mathcal{X}|$ denotes the cardinality of the set $\mathcal{X}$. As $N \to \infty$, the Glivenko–Cantelli theorem [30] implies that almost surely, $G_i^N(w) \to G_i(w)$ uniformly in w.

Let $D_i^N(t)$ denote the maximum delay of class i packets in the Nth system at time t. We assume that under all scheduling policies of interest, packets in a given class are served in the order of longest-delay-first.[9] In that case

$$D_i^N(t) = H_i^N(f_i^N(t)) + t \tag{9}$$

where $H_i^N(f) = \inf\{w : G_i(w) = f\}$. The first term in (9) corresponds to the maximum initial delay of the remaining packets; the second term corresponds to the aging of packets with time. It follows that in the limiting system, almost surely we have

$$D_i(t) = H_i(f_i(t)) + t \tag{10}$$

where $H_i(f_i(t))$ denotes the maximum delay of the remaining packets in the limiting system. If $G_i(w)$ is strictly increasing on $[D_i^{\min}, D_i^{\max}]$, then $H_i(x) = G_i^{-1}(w)$. Note that for finite N, the functions $G_i^N(f)$ and $D_i^N(t)$ are random quantities that depend on the initial delay distribution. However, in the limiting system, these quantities are deterministic.

In the Nth system, if the kth packet of class i is served at time t_k, then it receives a utility $U_i(D_i^N(t_k) + \frac{L}{R_i})$. The average utility per packet can be written as

$$U_{\text{avg}}^N = \frac{1}{N}\sum_{i=1}^{K}\sum_{k=1}^{N_i} U_i\left(D_i^N(t_k) + \frac{L}{R_i}\right).$$

As $N \to \infty$, we have $U_{\text{avg}}^N \to U_{\text{avg}}$, where

$$U_{\text{avg}} = \sum_{i=1}^{K}\int_0^{T_f} \alpha_i(t) R_i U_i\left(D_i(t)\right) dt. \tag{11}$$

C. Limiting Behavior of $\dot{U}R$ Scheduler

Next we characterize the limiting behavior of the $\dot{U}R$ scheduling rule for a TDM system. To simplify our analysis we focus on the case where $U_i(\cdot)$ is a decreasing concave function. For the fluid system, the scheduling decision is characterized by the parameter $\alpha_i(t)$ for each class i.

Let $S(t) = \{i : f_i(t) > 0\}$ be the set of nonempty classes, i.e., the classes with packets remaining to be sent at time t. Define $M_i(t) = |\dot{U}_i(D_i(t))|R_i$ to be the decision metric used by the scheduler for each class $i \in S(t)$. Among classes $i \in S(t)$, the $\dot{U}R$ scheduler transmits the Head of Line (HOL) packet of the class with the maximum value of $M_i(t)$ at each decision instant. Therefore, the $\dot{U}R$ policy has the following two properties:

Property 1: If $i \notin S(t)$, then $\alpha_i(t) = 0$.

Property 2: For $i \in S(t), \alpha_i(t) = 0$ if there exists $j \ne i$ such that $M_i(t) < M_j(t)$.

For $i \notin S(t)$, i.e., classes which are drained at time t, we will assume that $D_i(t) = D_i^{\min} + t$, which is a natural extension of (10). That is, the delay for class i formally continues to increase after all class i packets have been drained. Note that this does not affect any scheduling decisions or performance, but will be useful in Section V, where we formulate a fixed terminal-time optimal control problem.

As an example, consider a two-class system, i.e., $K = 2$. Assume that both classes have the same utility function, i.e., $U_1(D) = U_2(D) = U(D)$, and that the initial delays are uniformly distributed on $[0, 1]$, i.e., for $i = 1, 2$,

$$G_i(w) = \begin{cases} w, & 0 \le w \le 1, \\ 1, & w \ge 1. \end{cases} \tag{12}$$

In this case, (10) becomes

$$D_i(t) = f_i(t) + t \tag{13}$$

and therefore $\dot{D}_i(t) = -\frac{\alpha_i(t)R_i}{p_i} + 1$, with $D_i(0) = 1$ for $i = 1, 2$.

From Properties 1 and 2, we have

$$\alpha_1(t) = \begin{cases} 1, & \text{if } M_1(t) > M_2(t) \text{ or } f_2(t) = 0 \\ 0, & \text{if } M_1(t) < M_2(t) \text{ or } f_1(t) = 0 \end{cases}$$

and $\alpha_2(t) = 1 - \alpha_1(t)$. This specifies the scheduling rule except at those times t when $M_1(t) = M_2(t)$ and $f_i(t) > 0$ for $i = 1, 2$.

When multiple classes simultaneously have the maximum value of M_i, the fluid scheduler splits its resources among these. Let $Q(t)$ be the set of nonempty classes that have the maximum value of M_i, i.e.,

$$Q(t) = \{i \in S(t) : M_i(t) \ge M_j(t) \text{ for all } j \in S(t)\}.$$

The following theorem quantifies how resources are shared among these packets when $|Q(t)| \ge 2$.

[9] As noted in Section III.A, this will always be satisfied by both optimal policy and the $\dot{U}R$ policy when the utility functions are concave.

Theorem 1: Assume that for each $i = 1, \dots, K, U_i(\cdot)$ is concave. For any $t < T_f$ with $|Q(t)| \geq 2$, let $\{\alpha_i(t), i \in Q(t)\}$ be the solution to

$$\begin{aligned}\dot{M}_i(t) &= -\ddot{U}_i(D_i(t))\left(-\frac{\alpha_i(t)R_i}{p_i}\dot{H}_i(f_i(t)) + 1\right)R_i \\ &= \Lambda(t), \end{aligned} \tag{14}$$

where $\Lambda(t)$ is chosen to satisfy

$$\sum_{i\in Q(t)} \alpha_i(t) = 1. \tag{15}$$

If a feasible solution exists, i.e., $0 < \alpha_i < 1$ for all $i \in Q(t)$, then the scheduler spends α_i fraction of time serving class i packets.

In this theorem, (14) means that resources are shared at time t so that the decision metrics for all the classes in $Q(t)$ have the same derivative, where $\Lambda(t)$ is this common value; (15) implies that the value of $\Lambda(t)$ is such that all the resources are being used.

Proof: Consider some time t_0 for which there are two classes $i \neq j$ with $i, j \in Q(t_0)$; hence $M_i(t_0) = M_j(t_0)$. Let $\{\alpha_k^*(t_0)\}_{k\in Q(t)}$ denote the solution to (14) and (15) at t_0, and assume that these are all feasible. Suppose that the actual fraction of resources devoted to class i at t_0 is $\alpha_i(t_0) > \alpha_i^*(t_0) > 0$. Since the $\dot{U}R$ scheduler is nonidling it must satisfy (15); so, there must exist a class j such that $\alpha_j(t_0) < \alpha_j^*(t_0)$. Since $U(D)$ is concave, $\ddot{U}(D) < 0$ for all $D > 0$. Hence, $\dot{M}_i(t_0)$ is decreasing in $\alpha_i(t_0)$ from (14). Therefore, $\dot{M}_i(t_0) - \dot{M}_j(t_0) < 0$. Since $M_i(t_0) = M_j(t_0)$, we have $M_i(t_0^+) < M_j(t_0^+)$. From Property 2 we have $\alpha_i(t_0^+) = 0$. This violates the right continuity of $\alpha_i(t)$. A similar contradiction can be found if $\alpha_i(t_0) < \alpha_i^*(t_0)$. Therefore $\alpha_i^*(t_0)$ must be optimal. □

It can be shown that a unique solution to (14) and (15) always exists. Whether or not this solution satisfies $0 < \alpha_i < 1$ for all $i \in Q(t)$ depends on the choice of $U_i(D), H_i(f)$, and R_i. Given $H_i(f)$ and R_i, we define a set of utility functions $\{U_i(D)\}, i = 1, \cdots, K$, to be *regular* if a feasible solution to (14) and (15) exists for all t where $|Q(t)| \geq 2$. For example, with $K = 2, R_1 > R_2$, and uniform initial delays, $U_1(D) = U_2(D) = -D^\beta$, is regular for $\beta > 1$. In what follows, we will assume that $\{U_i(D)\}$ is regular unless stated otherwise.

For the two-class example described previously with uniform initial delay distribution, Theorem 1 implies that for any t such that $f_1(t), f_2(t) > 0$ and $\dot{U}(D_1(t))R_1 = \dot{U}(D_2(t))R_2$, the $\dot{U}R$ rule gives

$$\alpha_1(t) = \frac{\ddot{U}(D_1(t))R_1 - \ddot{U}(D_2(t))R_2 + \ddot{U}(D_2(t))R_2^2/p_2}{\ddot{U}(D_1(t))R_1^2/p_1 + \ddot{U}(D_2(t))R_2^2/p_2} \tag{16}$$

and $\alpha_2(t) = 1 - \alpha_1(t)$.

Define $t_i^{\text{in}} := \inf\{t : \alpha_i(t) > 0\}$ to be the time the server starts to serve class i packets. Likewise, define $t_i^{\text{out}} = \inf\{t : f_i(t) = 0\}$ to be the time at which all class i packets are drained.

Corollary 1: For regular $U_i(\cdot), \alpha_i(t) > 0$ for all $t \in (t_i^{\text{in}}, t_i^{\text{out}})$.

In other words, once the scheduler starts serving class i packets, it continues to serve this class until all class i packets are drained. This follows from Theorem 1, which implies that once class i joins the active set $Q(t)$, it remains in $Q(t)$ until time t_i^{out}. From Corollary 1, $t_i^{\text{out}} = \inf\{t > t_i^{\text{in}} : \alpha_i(t) = 0\}$.

The initiation and termination times for class i packets, $\{t_i^{\text{in}}\}_{i=1}^K$ and $\{t_i^{\text{out}}\}_{i=1}^K$, mark $2K$ events.[10] Let $t^1 \leq t^2 \leq \cdots \leq t^{2K}$ denote the ordered list of these times, i.e., $t^k = t_i^{\text{in}}$ or t_i^{out} for some i for each $k = 1, \dots, 2K$, where $t^1 = 0$, and $t^{2K} = T_f$.

Define the *upper envelope* of $\{M_i(t)\}_{i=1}^K$ to be

$$\bar{M}(t) = M_i(t), \quad i \in Q(t), \quad \text{for } t = [0, T_f). \tag{17}$$

This is the value of the decision metric for the classes that are being served. Notice that t_i^{in} and t_i^{out} satisfy $\bar{M}(t_i^{\text{in}}) = |\dot{U}_i(D_i^{\max} + t_i^{\text{in}})|R_i$ and $\bar{M}(t_i^{\text{out}}) = |\dot{U}_i(D_i^{\min} + t_i^{\text{out}})|R_i$. This is illustrated in Fig. 2, which shows an example of the upper envelope $\bar{M}(t)$ versus time. Also shown in the figure are the lines corresponding to the largest possible value of $|\dot{U}|R$ for each class (i.e., $|\dot{U}_i(D_i^{\max} + t)|R_i$) and the smallest possible value of $|\dot{U}|R$ for each class (i.e., $|\dot{U}_i(D_i^{\min} + t)|R_i$).[11] The intersections of these lines with $\bar{M}(t)$ mark the times t_i^{in} and $t_i^{\text{out}}, i = 1, 2$.

So far, we have characterized the $\dot{U}R$ rule given the decision metrics $\{M_i(t)\}_{i=1}^K$. Next, we determine how each decision metric $M_i(t)$ evolves with t. Recall that $Q(t)$ is the set of nonempty classes receiving service at time t. Let $\bar{Q}(t) = S(t) - Q(t)$ be the set of inactive classes, which still have packets remaining to be transmitted at time t. The decision metrics and the upper envelope can be computed via the iterative procedure in Algorithm 1. The quantities in step (2.a) of the algorithm can be computed directly from their definitions. In step (2.d), the two terms in the minimum are the smallest $t_i^{\text{in}} > t^k$ and the smallest $t_i^{\text{out}} > t^k$. Given $\bar{M}(t)$, the system behavior is completely determined. Namely, the event times $\{t^k\}$ are the intersections of $\bar{M}(t)$ with $|\dot{U}_i(D_i^{\max} + t)|R_i$ or $|\dot{U}_i(D_i^{\min} + t)|R_i$, for $i = 1, \cdots, K$. The evolution of the decision metrics and service allocations between successive event times is given by Theorem 1.

Algorithm 1: Iterative algorithm for calculating the decision metric trajectories.

```
1) Set $k = 1, t^1 = 0$.
2) While $t^k < T_f$ do:
        a) Calculate $f_i(t^k)$ and $M_i(t^k)$ and update
           $S(t^k), Q(t^k)$ and $\bar{Q}(t^k)$;
        b) Set $\alpha_i(t^k) = 0$ for $i \notin S(t^k)$;
        c) If $Q(t) = \{i\}$, set $\alpha_i(t) = 1$ and $\alpha_j(t) = 0$
           for all $j \notin Q(t)$ for $t \in (t^k, t^{k+1})$.
              else if $|Q(t)| \geq 2$, calculate $\alpha_i(t)$ for
              $i \in Q(t)$ and $t \in (t^k, t^{k+1})$ from Theorem 1;
        d) Evaluate $\bar{M}(t)$ from (14) and (17) for
           $t \in (t^k, t^{k+1})$, and compute
           $t^{k+1} = \min[\inf(t : M_j(t) = \bar{M}(t), j \in \bar{Q}(t)),$
                              $\inf(t : f_i(t) = 0, \forall i \in Q(t))]$;
        e) Set $k = k + 1$.
End While.
```

[10] It is possible that some of these events coincide. In that case, we can order them arbitrarily.

[11] In Fig. 2 the curves are lines; this corresponds to quadratic utilities.

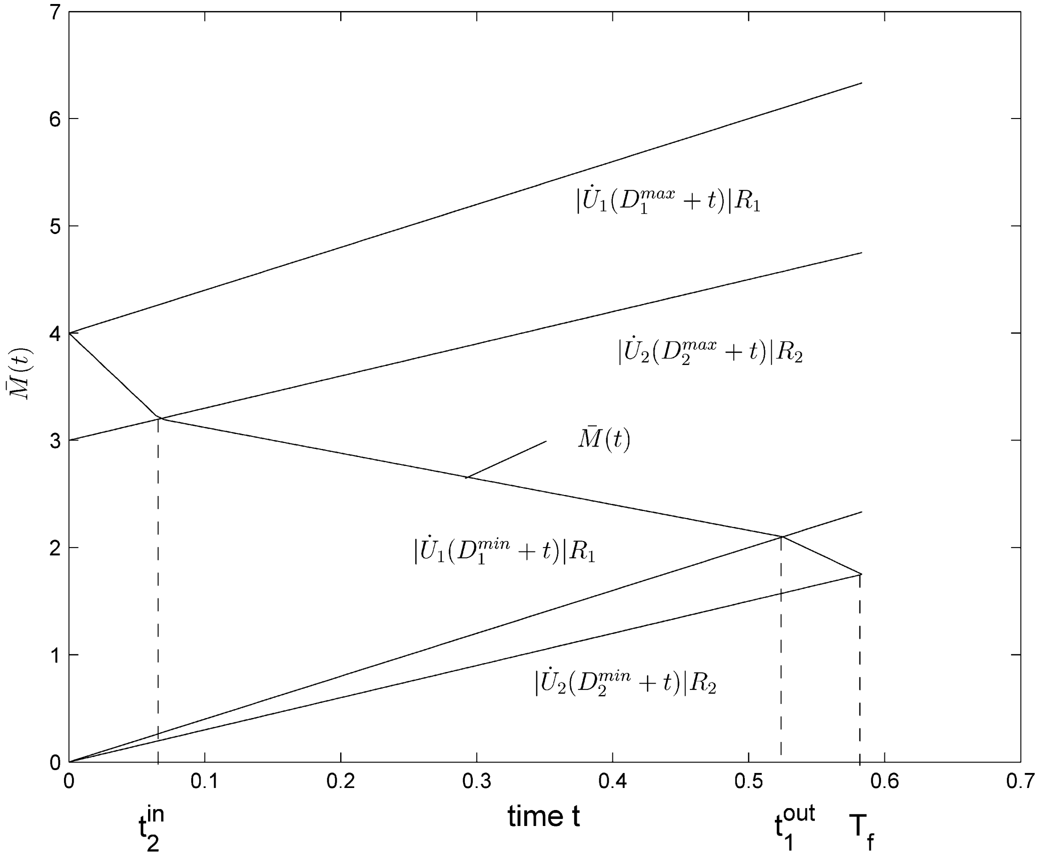


Fig. 2. An example of the time evolution of $\bar{M}(t)$, $|\dot{U}_i(D_i^{\max}+t)|R_i$, and $|\dot{U}_i(D_i^{\min}+t)|R_i$ for a two-class system. The intersections of $\bar{M}(t)$ with the other curves correspond to the times $t = 0, t_2^{\rm in}, t_1^{\rm out}, T_f$.

For the two-class example with uniform initial delay distribution the preceding procedure gives: i) $\alpha_1(t) = 1$ for $t \in [0, t_2^{\rm in})$, ii) $\alpha_1(t)$ is given by (16) for $t \in [t_2^{\rm in}, t_1^{\rm out})$, and iii) $\alpha_1(t) = 0$ for $t \in [t_1^{\rm out}, T_f]$, where $t_2^{\rm in}$ satisfies $\dot{U}(1+(1-R_1/p_1)t_2^{\rm in})R_1 = \dot{U}(1+t_2^{\rm in})R_2$. Here, $\alpha_2(t) = 1-\alpha_1(t)$. Also, from step (2.d) of the iteration, if $0 < t_2^{\rm in} < \frac{p_1}{R_1}$, then $t_1^{\rm out} > t_2^{\rm in}$; otherwise, $t_1^{\rm out} = t_2^{\rm in} = \frac{p_1}{R_1}$.

D. *Numerical Example* $U(D) = -D^\beta$

We illustrate the preceding analysis for two classes, each with utility function $U(D) = -D^\beta$, where $\beta > 1$. Note that $U(D)$ is concave. We also assume that $R_1 > R_2$. Each class i has probability p_i.

From Theorem 1, during the period when the two classes are simultaneously served we have $M_1(t) = M_2(t)$, and $\dot{M}_1(t) = \dot{M}_2(t)$. Substituting $M_i(t) = |\dot{U}_i(D_i(t))|R_i$ and $U(D) = -D^\beta$ into these expressions yields

$$D_1^{\beta-1}(t)R_1 = D_2^{\beta-1}(t)R_2 \tag{18}$$

and

$$D_1^{\beta-2}(t)\left(-\frac{\alpha_1 R_1}{p_1}+1\right)R_1 = D_2^{\beta-2}(t)\left(-\frac{\alpha_2 R_2}{p_2}+1\right)R_2. \tag{19}$$

Combining these two relations, it is straightforward to show that

$$\left(-\frac{\alpha_1 R_1}{p_1}+1\right)^{\beta-1}R_1 = \left(-\frac{\alpha_2 R_2}{p_2}+1\right)^{\beta-1}R_2 \tag{20}$$

independent of the current time t and delay $D_i(t)$. Solving (20) for $\alpha_1(t) = 1-\alpha_2(t)$ gives

$$\alpha_1(t) = \frac{\left(\frac{R_2}{R_1}\right)^{\frac{1}{\beta-1}}\frac{R_2}{p_2} - \left(\frac{R_2}{R_1}\right)^{\frac{1}{\beta-1}} + 1}{\frac{R_1}{p_1} + \left(\frac{R_2}{R_1}\right)^{\frac{1}{\beta-1}}\frac{R_2}{p_2}} \tag{21}$$

which is independent of t, i.e., the server is "statically" split between the two classes. As $\beta \to \infty, \alpha_1$ increases and approaches $\frac{R_2/p_2}{R_1/p_1+R_2/p_2}$. Therefore, in this limit $\frac{\alpha_1}{\alpha_2} \to \frac{R_2/p_2}{R_1/p_1}$ and the rate at which each class is drained is the same, i.e.

$$-\dot{f}_i(t) = \frac{\alpha_i(t)R_i}{p_i} = \frac{R_1 R_2}{R_1 p_2 + R_2 p_1}, \quad i = 1, 2.$$

If we further assume a uniform initial delay distribution for both classes given by (12), and that $p_1 = p_2 = 1/2$, then $t_2^{\rm in}$ satisfies

$$\left(1 - 2R_1 t_2^{\rm in} + t_2^{\rm in}\right)^{\beta-1} R_1 = \left(1 + t_2^{\rm in}\right)^{\beta-1} R_2 \tag{22}$$

and so

$$t_2^{\rm in} = \frac{1-\left(\frac{R_2}{R_1}\right)^{\frac{1}{\beta-1}}}{\left(\frac{R_2}{R_1}\right)^{\frac{1}{\beta-1}} + 2R_1 - 1}. \tag{23}$$

As $\beta \to \infty, t_2^{\rm in} \to 0$. Hence, the $\dot{U}R$ scheduler becomes a round-robin scheduler, i.e., it drains both classes at the same rate starting from $t = 0$.

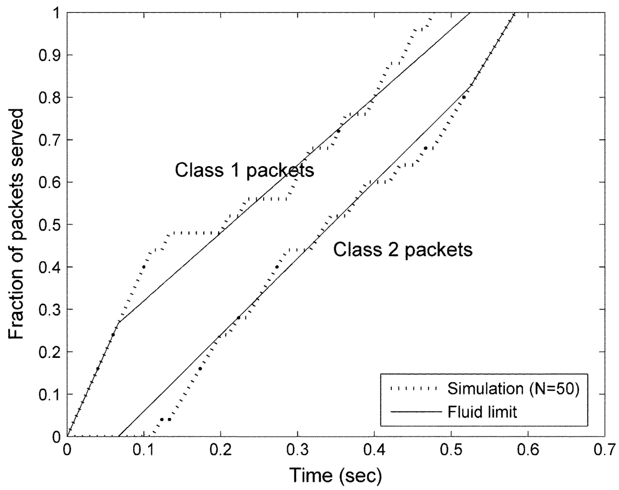


Fig. 3. Fraction of packets served over time with $U(D) = -D^2$ and uniform initial delays.

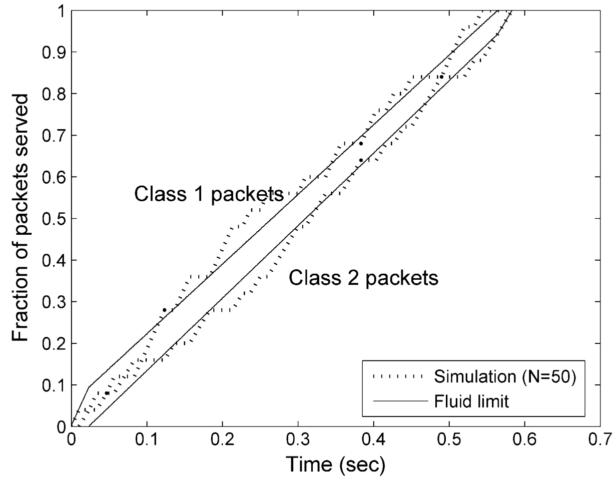


Fig. 4. The fraction of packets served over time with $U(D) = -D^4$ and uniform delays.

Fig. 3 shows the fraction of class 1 and class 2 packets served up to time t with $U(D) = -D^2$ and $R_1 = 2$ and $R_2 = 1.5$. In this case, $\alpha_1(t) = \frac{R_1 - R_2 + 2R_2^2}{2R_1^2 + 2R_2^2}$. According to Property 2, the scheduler first serves class 1 packets up to time $t_2^{\text{in}} = \frac{R_1 - R_2}{2R_1^2 - R_1 + R_2} = 1/15$. Then the scheduler drains the two classes simultaneously with $\alpha_1 = 2/5$ and $\alpha_2 = 3/5$. At time $t_1^{\text{out}} = \frac{R_2(R_1 + R_2)}{R_1(R_1 - R_2 + 2R_2^2)} = 21/40$, the scheduler finishes serving all the class 1 packets and starts to serve only class 2 packets until $T_f = 7/12$ when all packets are drained. This is represented by the solid lines, where the slope of each line at time t is $\frac{\alpha_i(t)R_i}{p_i}$. The dashed lines are from a sample run with $N = 50$ packets.

A similar plot is shown in Fig. 4 with $U(D) = -D^4$. The scheduler again initially serves only class 1 packets, then statically splits its service rate between the two classes until all class 1 packets are drained, and subsequently serves only class 2 packets. Comparing with Fig. 3, the resource-sharing period in this case starts earlier and lasts longer.

To study how well the asymptotic results predict the performance of a finite system, we simulated the $\dot{U}R$ scheduler for different numbers of packets, N. The simulation results are shown in Fig. 5, which shows sample values of the average utility per packet for different values of N. As expected, the variance of the utility decreases, and the utilities approach the fluid limit as N increases.

Next we compare the $\dot{U}R$ scheduler with the "Maximum Rate (Max R)" scheduler, which always schedules a packet from a class with the highest transmission rate. Within each class both schedulers transmit packets in the order of largest-delay-first. The Max R rule maximizes the aggregate data rate at any time t, but at the expense of increasing the delay variance. Fig. 6 shows the aggregated utility vs. time under both the Max R and $\dot{U}R$ policies. The $\dot{U}R$ rule generates greater average utility over all packets than the Max R rule. Initially, the Max R scheduler generates higher utility since it serves only class 1 packets at the higher rate. The Max R utility then drops below the $\dot{U}R$ utility once the longer delays experienced by class 2 packets dominate.

IV. Extension to Continuous Rate Distribution

In the previous section, we characterized the behavior of the $\dot{U}R$ rule when each packet is in one of a finite number of rate classes. We now relax this assumption, and assume that the rate for each packet is selected according to a continuous distribution, $p(r)$. The rates are still i.i.d. across packets and are independent of the initial packet delays. For a continuous rate distribution, the limiting system can be viewed as having an infinite number of rate classes.

For simplicity, we assume that all packets have the same utility function $U(D)$ and that the rate *pdf* $p(r)$ has bounded support, i.e., there exist $R_{\min} > 0$ and $R_{\max} < \infty$ such that $p(r) = 0$ for any $r < R_{\min}$ or $r > R_{\max}$. Each packet has a feasible rate r and an initial delay w, which are chosen independently with *pdf*'s $p(r)$ and $g(w)$, respectively. We still assume that $g(w) > 0$ is continuous and has compact support on $[D^{\min}, D^{\max}]$.

Once again we consider a fluid limit in which $N \to \infty$ and $L \to 0$ with $NL = 1$. With probability one, the total time required to drain the limiting system with any nonidling scheduling rule is now given by

$$T_f = \int_{R_{\min}}^{R_{\max}} \frac{p(r)}{r} dr. \tag{24}$$

Define $f^N(t, r, \delta r)$ to be the remaining fraction of packets in the Nth system with rate $\tilde{r} \in (r, r + \delta r)$, i.e.

$$f^N(t, r, \delta r) = \frac{\mathcal{N}^N(t, r, \delta r)}{\mathcal{N}^N(0, r, \delta r)}$$

where $\mathcal{N}^N(t, r, \delta r)$ denotes the remaining number of packets with rate $\tilde{r} \in (r, r + \delta r)$ at any time t. Let $f(t, r) = \lim_{\delta r \to 0} \lim_{N \to \infty} f^N(t, r, \delta r)$; this can be viewed as the density of remaining packets at time t. Following similar arguments as in Section III-B, this can be shown to satisfy

$$f(t, r) = 1 - \int_0^t \frac{\alpha(\tau, r) r}{p(r)} d\tau \tag{25}$$

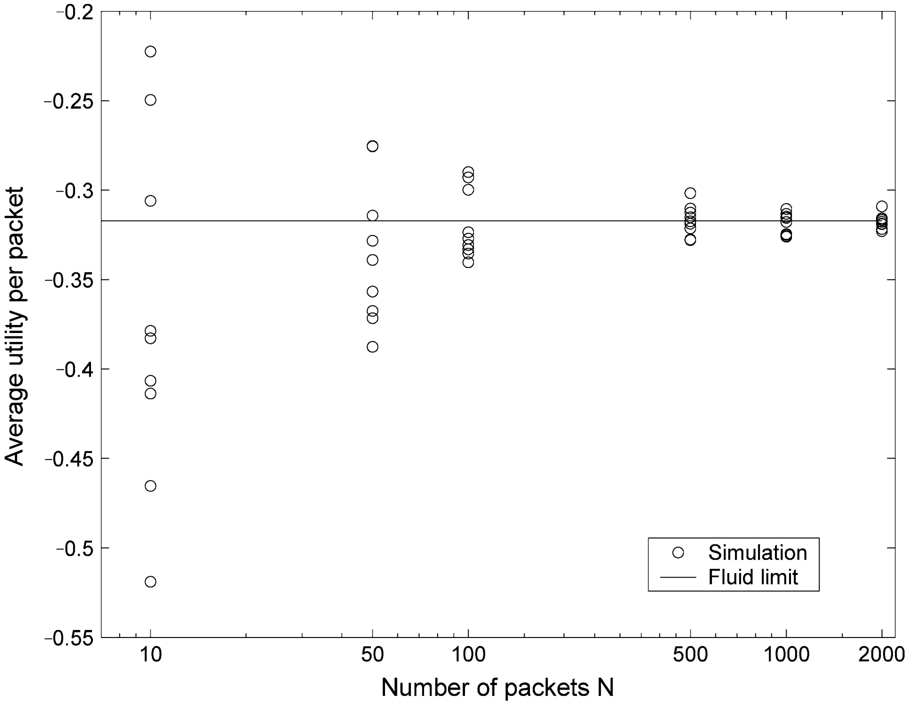


Fig. 5. The average utility per packet for different realizations of finite systems with N packets.

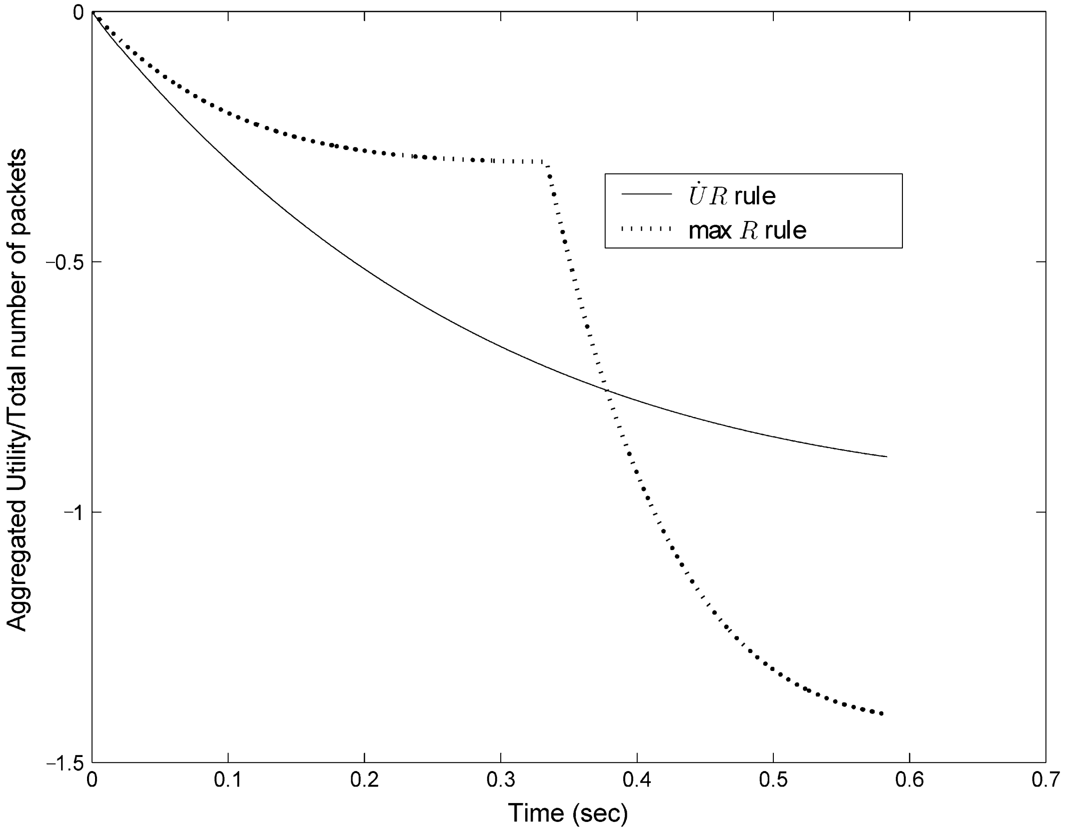


Fig. 6. The normalized aggregate utility over time for the Max R and $\dot{U}R$ rules.

where $\alpha(\tau, r)$ can be interpreted as the density of resources devoted to packets with rate r at time t. In analogy with the finite-class scenario, for any $t \in [0, T_f)$, this must satisfy

$$\int_{R_{\min}}^{R_{\max}} \alpha(t,r)dr = 1 \quad \text{and} \quad \alpha(t,r) \geq 0.$$

The longest delay for packets with rate r is therefore given by

$$D(t,r) = H(f(t,r)) + t$$

where $H(x) = G^{-1}(w)$ for strictly decreasing $G(w)$, and $G(w) = \int_0^w g(x)dx$.

Again, let $S(t) := \{r : f(t,r) > 0\}$ denote the set of nonempty rates at time t. We define $Q(t) = \{r : \alpha(t,r) > 0\}$ to be the set of rates corresponding to users that are actively served at time t. For each $t \in [0, T_f)$, let $\bar{M}(t) := \sup_r |\dot{U}(D(t,r))| r$ be the largest metric at that time. Notice that $\bar{M}(0) = \dot{U}(D^{\max})R_{\max}$.

As in the finite-rate model, there are event times which correspond to the start and end of service for packets with the same rate class. However, for the continuous-rate model, there are an uncountable number of such events corresponding to every possible rate r, and the active set $Q(t)$ is no longer finite. Under our simplified assumptions that all packets have the same utility function and the same initial delay distribution, the active set $Q(t)$ is a closed interval. The next lemma specifies $Q(t_0)$ for $t_0 \in [0, T_f)$ given $\bar{M}(t), 0 < t < t_0$.

Lemma 1: Let r_l satisfy $\bar{M}(t_0) =\mid \dot{U}(D^{\max}+t_0) \mid r_l$ and r_u satisfy $\bar{M}(t_0) =\mid \dot{U}(D^{\min}+t_0) \mid r_u$. At time $t_0 \in [0, T_f)$, the active set $Q(t) = [r_{\min}, r_{\max})$, where $r_{\min} = \max(R_{\min}, r_l)$ and $r_{\max} = \min(R_{\max}, r_u)$.

Proof: By definition, t_0 is the service initiation time for packets with rate r_l, hence $r_l \in Q(t)$ if $r_l > R_{\min}$. Likewise, t_0 is also the service termination time for packets with rate r_u, so $r_u \in Q(t)$ if $r_u < R_{\max}$. Since $\mid \dot{U}(D^{\max}+t_0) \mid r$ is a strictly increasing function of r, service of any packets with $r < r_l$ cannot have been initiated for any $t \le t_0$. Similarly, since $\mid \dot{U}(D^{\min}+t_0) \mid r$ is also strictly increasing in r, service to any packets with $r < r_u$ cannot have been terminated for any $t \le t_0$. Therefore $Q(t) = [r_{\min}, r_{\max})$. □

Lemma 1 implies that the starting service time for a packet with rate r is earlier than that for a packet with rate $r' < r$. Furthermore, packets with a higher rate are always depleted earlier than those with a lower rate. We emphasize that this is based on the assumptions that all packets have the same utility function, the initial delays are chosen from the same distribution, and the delays are independent of the transmission rates.

The next proposition specifies how $\bar{M}(t)$ evolves with time.

Proposition 2: Given $\bar{M}(t), 0 < t < t_0$, let $\alpha(r, t_0)$ be the solution to

$$\begin{aligned}&\frac{d\bar{M}(t)}{dt}\Big|_{t\to t_0^+}\\ &= -\ddot{U}(D(t_0,r))\left(\left(-\frac{\alpha(t_0,r)r}{p(r)}\right)\dot{H}(f(t_0,r))+1\right)r\\ &= \Lambda(t_0)\end{aligned} \tag{26}$$

where $\Lambda(t_0)$ is chosen such that

$$\int_{Q(t_0)} \alpha(r,t_0)dr = 1, \quad \text{for all } t_0 \in [0,T_f). \tag{27}$$

If a feasible solution exists, i.e., $\alpha(r,t_0) > 0$ for all $r \in Q(t_0)$, then the scheduler serves associated packets with rate $\alpha(r,t_0)r$.

The proof is very similar to that of Theorem 1, so we omit it. Note that the active set $Q(t_0)$ in (27) is in turn determined by Lemma 1. Solving the differential equation (26) in Proposition 2, we can derive the trajectory $\bar{M}(t)$ for $t \in [0, T_f)$. The utility per packet (for any scheduler) is given by

$$U_{\text{avg}} = \int_{R_{\min}}^{R_{\max}} \int_0^{T_f} \alpha(t,r) r U(D(t,r))dt. \tag{28}$$

A. Numerical Example

We give a numerical example to illustrate the preceding analysis. Assume a rate *pdf* $p(r) = Kr^{-3/2}$ for $r \in [1, 10]$. This corresponds to the situation in which the transmission rate is proportional to the received power, which is determined from distance-based attenuation with a path-loss exponent of four, and the users are uniformly distributed within the unit circle. Let the initial packet delay be uniformly distributed on the interval $[0, 1]$. The utility function for all packets is $U(D) = -\frac{1}{2}D^2$.

In this case, (26) becomes

$$\left(-\frac{\alpha(t_0,r)r}{p(r)}+1\right)r = \Lambda(t_0)$$

and combining with (27) gives

$$\frac{d\bar{M}(t)}{dt}\Big|_{t\to t_0^+} = \Lambda(t_0) = \frac{\int_{r_{\min}}^{r_{\max}} p(r)/r dr - 1}{\int_{r_{\min}}^{r_{\max}} p(r)/r^2 dr} \tag{29}$$

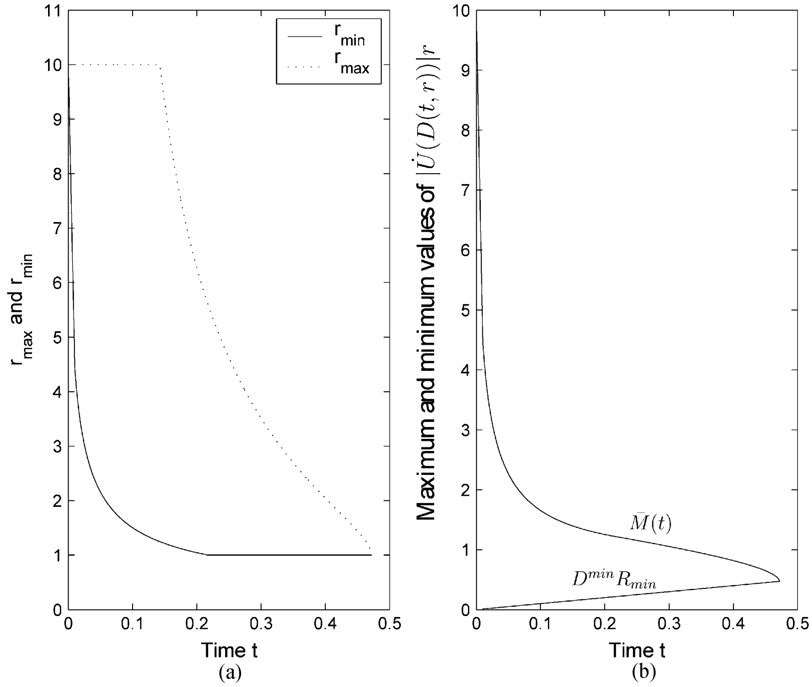
where, from Lemma 1, $r_{\min} = \max(1, \frac{\bar{M}(t_0)}{1+t_0})$ and $r_{\max} = \min(10, \frac{\bar{M}(t_0)}{t_0})$.

Using (29) to calculate the upper envelope $\bar{M}(t)$ yields the curves shown in Fig. 7. Fig. 7(a) shows how $r_{\min}$ and $r_{\max}$ change over time.[12] Initially, $r_{\min} = r_{\max} = R_{\max} = 10$. As classes join service, $r_{\min}$ decreases while $r_{\max}$ stays fixed. At $t \approx 0.134$, all packets with rate $r = R_{\max} = 10$ are drained, and then $r_{\max}$ starts to decrease. When $r_{\min}$ reaches $R_{\min} = 1$, all rates have become active. Subsequently, $r_{\min}$ stays at the minimum while $r_{\max}$ keeps decreasing until the terminal time T_f when all packets are drained and $r_{\min} = r_{\max} = R_{\min} = 1$. Fig. 7(b) shows how $\bar{M}(t)$ evolves with time. The minimum metric $|\dot{U}(D^{\min}+t)|R_{\min} = t$ is also shown. At $t = T_f$, the two curves merge, which signifies that all packets are drained.

V. Optimality of $\dot{U}R$ Policy With TDM Capacity Regions

In this section, we discuss an optimal scheduling problem for the fluid system with a TDM capacity region. For simplicity, we consider a $K = 2$ class system with transmission rates R_i and concave decreasing utility functions $U_i(D)$, for $i = 1, 2$. The probability a packet is assigned to class i is given by p_i. We again assume that the initial delay for class i packets is distributed on the interval $[D_i^{\min}, D_i^{\max}]$ according to the c.d.f. $G_i(w)$, with a well-defined inverse $H_i(x)$. Without loss of generality, assume that $|\dot{U}_1(D_1(0))|R_1 \ge |\dot{U}_2(D_2(0))|R_2$ so that $t_1^{\text{in}} = 0$, i.e., the scheduler always begins by serving class

[12] Note that $r_{\min}$ and $r_{\max}$ are parameters of the algorithm and correspond to the active set, whereas $R_{\min}$ and $R_{\max}$ correspond to the limits of the rate density.



(a)

(b)

Fig. 7. (a) Range of active rates from $r_{\min}$ to $r_{\max}$; (b) Decision metric envelope $\bar{M}(t)$ versus t.

1 packets. Characterizing a scheduling policy is equivalent to specifying the functions $\alpha_1(t)$ and $\alpha_2(t)$ for all $t \in [0, T_f)$. We want to choose these to maximize the total utility derived.

Formally, this problem can be written as

Problem OC:

$$\min_{\alpha_1(t),\alpha_2(t)} \quad \int_0^{T_f} \left[-\sum_{i=1}^{2} \alpha_i(t) R_i U_i(H_i(f_i(t)) + t) \right] dt \tag{30}$$

$$\text{subject to :} \quad \dot{f}_i(t) = -\frac{\alpha_i(t)}{p_i} R_i, \quad i = 1, 2 \tag{31}$$

$$f_i(0) = 1, \quad i = 1, 2 \text{ and } f_1(T_f) = 0 \tag{32}$$

$$\alpha_1(t) + \alpha_2(t) = 1 \tag{33}$$

$$\alpha_i(t) \geq 0, \quad i = 1, 2. \tag{34}$$

This can be viewed as a continuous-time optimal control problem [32] with a fixed terminal time T_f, where the state is $\mathbf{f}(t) = (f_1(t), f_2(t))$, and $\boldsymbol{\alpha}(t) = (\alpha_1(t), \alpha_2(t))$ is the control variable. Here (31) represents the system dynamics, and (32) gives initial and final boundary conditions for the state. The final state $(f_1(T_f), f_2(T_f))$ is restricted to be on the line $f_1(T_f) = 0$. Any admissible control $\boldsymbol{\alpha}(t)$ also results in $f_2(T_f) = 0$. However, we do not need to explicitly state this boundary condition. If we are given $f_1(t)$, then we can compute $f_2(t)$ and in particular, $f_1(T_f) = 0$ implies $f_2(T_f) = 0$. Hence the latter constraint is not independent. Furthermore, we require $\boldsymbol{\alpha}(t)$ to be a right-continuous function. It can be shown that an optimal control satisfying this assumption exists.[13]

If all the packets in class i are emptied at time $\hat{t} < T_f$, then for all $t > \hat{t}$, we have that $\alpha_i(t) = 0$ and $f_i(t) = 0$. To see that this must hold in the preceding formulation, note that since $f_i(\hat{t}) = 0$ and $f_i(T_f) = 0, \dot{f}_i(t) = 0$ for $\hat{t} < t < T_f$, and from (31) and (34), $\alpha_i(t) = 0$ for $\hat{t} < t < T_f$.

The solution to this problem can be characterized using the Pontryagin minimum principle [32]. We first define the Hamiltonian for this problem, which is given by

$$\mathcal{H}(\mathbf{f}(t), \boldsymbol{\alpha}(t), \mathbf{q}(t)) = -\sum_{i=1}^{2} \alpha_i(t) R_i \left[U_i(D_i(t)) + \frac{q_i(t)}{p_i} \right]$$

where $\mathbf{q}(t) = (q_1(t), q_2(t))$ is the co-state or Lagrange multiplier, and $D_i(t) = H_i(f_i(t)) + t$. Let $\boldsymbol{\alpha}^*(t)$ be an optimal control and $\mathbf{D}^*(t)$ the corresponding optimal state trajectory. According to the Pontryagin minimum principle, there exists a $\mathbf{q}^*(t)$ such that

$$\dot{\mathbf{q}}^*(t) = -\nabla_{\mathbf{f}} \mathcal{H}(\mathbf{f}^*(t), \boldsymbol{\alpha}^*(t), \mathbf{q}^*(t)) \tag{35}$$

and

$$\mathcal{H}(\mathbf{f}^*(t), \boldsymbol{\alpha}^*(t), \mathbf{q}^*(t)) \leq \mathcal{H}(\mathbf{f}^*(t), \boldsymbol{\alpha}(t), \mathbf{q}^*(t)) \tag{36}$$

for all admissible controls $\boldsymbol{\alpha}(t)$.

For this problem, the co-state (35) are

$$\dot{q}_i(t) = \alpha_i(t) R_i \dot{U}_i(D_i(t)) \dot{H}_i(f_i(t)), i = 1, 2.$$

Furthermore, the final state conditions dictate that $q_2(T_f) = 0$ [33]. Let $A_i(t) = R_i[U_i(D_i(t)) + \frac{q_i(t)}{p_i}]$ for $i = 1, 2$. Then the Hamiltonian can be written as

$$\mathcal{H}(\mathbf{f}(t), \boldsymbol{\alpha}(t), \mathbf{p}(t)) = -A_1(t)\alpha_1(t) - A_2(t)\alpha_2(t),$$

[13] Since the dynamics are linear and the feasible control set is compact, there exists an absolutely continuous solution, (f_1^*, f_2^*), to this problem. From the dynamics, it can be seen that f_i^* must be nonincreasing. Letting $\alpha_i(t)$ be its right derivative gives a right-continuous optimal control.

which is linear in $\alpha_i(t)$. Hence, to satisfy (36), it follows that

$$\alpha_1^*(t) = \begin{cases} 1, & \text{if } A_1(t) > A_2(t) \\ 0, & \text{if } A_1(t) < A_2(t) \end{cases} \tag{37}$$

and $\alpha_2^*(t) = 1 - \alpha_1^*(t)$. Let $\Delta(t) = A_1(t) - A_2(t)$. If $\Delta(t) = 0$, then the problem is said to be *singular* at time t. This means that (36) alone does not specify the optimal control. A *singular interval* $[t_1, t_2]$ means that the problem is singular for all t in $[t_1, t_2]$, i.e., $\Delta(t) = 0$ for all $t \in [t_1, t_2]$.

Lemma 2: During any singular interval, the optimal control must satisfy (16).

Proof: Notice that

$$\begin{aligned}\dot{A}_i(t) &= R_i \left[\dot{U}_i(D_i(t)) \dot{D}_i(t) + \frac{\dot{q}_i(t)}{p_i} \right] \\ &= R_i \left\{ \dot{U}_i(D_i(t)) \left[-\dot{H}_i(f_i(t)) \frac{\alpha_i(t)}{p_i} R_i + 1 \right] \right. \\ &\quad \left. + \frac{\alpha_i(t)}{p_i} R_i \dot{U}_i(D_i(t)) \dot{H}_i(f_i(t)) \right\} \\ &= R_i \dot{U}_i(D_i(t)) \qquad (38) \\ &= -M_i(t) \qquad (39)\end{aligned}$$

which does not depend on $\alpha_i(t)$. Furthermore, for all $t \in [t_1, t_2]$, it must be that $\dot{\Delta}(t) = 0$. Therefore, $\dot{A}_1(t) = \dot{A}_2(t)$, i.e., $R_1 \dot{U}_1(D_1(t)) = R_2 \dot{U}_2(D_2(t))$. This corresponds to the choice of $\alpha_1(t)$ in (16). □

Therefore, the sign of $\Delta(t)$ determines the optimal control at time t. Notice that $\Delta(t)$ is continuous and differentiable since both $A_1(t)$ and $A_2(t)$ are continuous and differentiable.

Lemma 2 implies that during any singular interval, the optimal scheduling policy behaves like the $\dot{U}R$ rule. Recall from Section 3.3 that the $\dot{U}R$ starts serving class 1 packets up to t_2^{in}, then serves both classes simultaneously for[14] $t_2^{\text{in}} \le t \le \min\{t_2^{\text{out}}, t_1^{\text{out}}\}$, and finally devotes service to the remaining class until $t = T_f$.[15] To show that the $\dot{U}R$ rule is optimal for all $t \in [0, T_f)$, we still need to show that i) $\Delta(t)$ is unique; ii) $\Delta(t) > 0$ for $t \in [0, \hat{t}_2^{\text{in}}), \Delta(t) = 0$ for $t \in [\hat{t}_2^{\text{in}}, \min\{\hat{t}_1^{\text{out}}, \hat{t}_2^{\text{out}}\})$, and $\Delta(t) < 0$ for $t \in [\hat{t}_1^{\text{out}}, T_f)$ (if $\hat{t}_1^{\text{out}} < \hat{t}_2^{\text{out}}$) or $\Delta(t) > 0$ for $t \in [\hat{t}_2^{\text{out}}, T_f)$ (if $\hat{t}_2^{\text{out}} < \hat{t}_1^{\text{out}}$); and iii) $t_2^{\text{in}} = \hat{t}_2^{\text{in}}, t_1^{\text{out}} = \hat{t}_1^{\text{out}}$ and $t_2^{\text{out}} = \hat{t}_2^{\text{out}}$.

In the following, we assume that $U_1(D)$ and $U_2(D)$ are decreasing, strictly concave in D, and that they are regular (see Section III-C) for the given delay distributions and rates. We first show in Lemma 3 that for such utility functions, if $\Delta(t)$ is nonincreasing on an interval where it is strictly positive, then it must be strictly decreasing on this interval. Next, in Lemma 4, we show that if $\Delta(t)$ is nonincreasing, then the $\dot{U}R$ rule must be optimal. Finally, in Theorem 2, we give a condition on the utility functions under which the $\dot{U}R$ rule is optimal. The proofs are given in the Appendices.

Lemma 3: Let $I = [a, b)$ be a half-open interval such that $\Delta(t) > 0$ for all $t \in I$. If $\Delta(t)$ is nonincreasing, i.e., $\dot{\Delta}(t) \le 0$ for all $t \in I$, then for regular utility functions, $\dot{\Delta}(t) < 0$ for all $t \in I$.

Lemma 4: For regular utility functions, if $\dot{\Delta}(t) \le 0$ for all $t \in [0, T_f]$, then the $\dot{U}R$ rule is optimal.

Theorem 2: Assume that the utility functions satisfy the following condition for all $t_0 > 0$:

If $R_1 \dot{U}_1(D_1(t_0)) = R_2 \dot{U}_1(D_2(t_0))$, then for all $s > 0$,

$$\text{i)} \quad R_1 \dot{U}_1 \left(H_1 \left(f_1(t_0) - \frac{R_1}{p_1} s \right) + t_0 + s \right) > R_2 \dot{U}_2(D_2(t_0) + s);$$

and

$$\text{ii)} \quad R_1 \dot{U}_1(D_1(t_0) + s) < R_2 \dot{U}_2 \left(H_2 \left(f_2(t_0) - \frac{R_2}{p_2} s \right) + t_0 + s \right).$$

Then the $\dot{U}R$ rule is optimal.

Recall, $f_i(t_0) = G_i(D_i(t_0) - t_0)$ is the fraction of class i packets remaining at time t_0, where $G_i(w)$ is the c.d.f. of the initial delay distribution for class i. The left-hand (right-hand) side of condition i) is the value of $M_1(t_0 + s)(M_2(t_0 + s))$ if the scheduler serves only class 1 packets from time t_0 to $t_0 + s$. Condition ii) is the analogous relation if the scheduler serves only class 2 packets.

Corollary 2: With a uniform initial delay distribution for each class,the $\dot{U}R$ rule is optimal in the following cases:

1) $U(D) = -D^\beta$ with $\beta > 1$ and $R_1 > R_2 > 0$.
2) $U(D) = 1 - e^{kD}$ where $k > 0$ is a constant and $R_1 > R_2 > 0$.
3) $U(D)$ is concave and $R_1 > R_2 > 1$.

VI. Optimality for Non-TDM Capacity Regions

In this section, we consider the optimality of the $\dot{U}R$ rule for a more general two-user capacity region $\mathcal{C}$ that is a compact, convex and coordinate convex[16] subset of $\mathbb{R}_+^2$. For an arbitrary capacity region, we define $\delta\mathcal{C}$ to be the set of Pareto dominate rates, i.e., $\mathbf{r} \in \delta\mathcal{C}$ if and only if $\mathbf{r} \in \mathcal{C}$ and there is no other $\mathbf{r}' \in \mathcal{C}$ such that $\mathbf{r}' \ge \mathbf{r}$. (All vector inequalities are component-wise.) We say that $\mathcal{C}$ has a strictly convex boundary if for any pair $\mathbf{r}, \mathbf{r}' \in \delta\mathcal{C}$, $\alpha\mathbf{r} + (1 - \alpha)\mathbf{r}' \notin \delta\mathcal{C}$ for any $\alpha \in (0, 1)$. One example of a capacity region $\mathcal{C}$ with a strictly convex boundary is the achievable rate region for a Gaussian broadcast channel. A rate vector $\mathbf{r} = (r_1, r_2)$ is defined to be in the *interior* of $\delta\mathcal{C}$ if $\mathbf{r} \in \delta C$ and $\mathbf{r} > \mathbf{0}$, i.e., both users receive a positive rate.

With such a capacity region, the $\dot{U}R$ scheduling policy selects a rate vector $\mathbf{r}(t) = (r_1(t), r_2(t))$ at each time t such that

$$\mathbf{r}(t) = \arg\max_{\mathbf{r} \in \mathcal{C}} \sum_{i=1}^{2} |\dot{U}_i(D_i(t))| r_i. \tag{40}$$

Note that with the preceding assumptions, this optimization problem always has a solution $\mathbf{r} \in \delta\mathcal{C}$, and if $\mathcal{C}$ has a strictly

[14] Either of these intervals may have measure zero, e.g., when $t_2^{\text{in}} = t_1^{\text{out}}$.

[15] Which class remains depends on which of t_1^{out} and t_2^{out} is smaller. This in turn depends on the utilities and delay distributions. If $U_1(D) = U_2(D)$ and $D_1^{\min} = D_2^{\min}$, then for $R_1 > R_2$, $t_1^{\text{out}} \le t_2^{\text{out}}$, and so class 2 is the remaining class.

[16] A set $\mathcal{X} \subset \mathbb{R}_+^n$ is said to be *coordinate convex* if $\mathbf{x} \in \mathcal{X}$ implies that $\mathbf{y} \in \mathcal{X}$ for all $\mathbf{y}$ such that $\mathbf{0} \le \mathbf{y} \le \mathbf{x}$.

convex boundary, then the solution is unique. For a given capacity region, $\mathcal{C}$, at each time t, the solution to (40) depends only on the ratio $V(t) \equiv \frac{\dot{U}_1(D_1(t))}{\dot{U}_2(D_2(t))}$. If $\mathcal{C}$ has a strictly convex boundary, then given any point $\hat{\mathbf{r}}$ in the interior of δC there is a unique value of the ratio $V(t)$ for which $\hat{\mathbf{r}}$ is the solution to (40).

The corresponding optimal control problem in this setting is given by

$$\begin{aligned} \min_{r_1(t), r_2(t)} \quad & \int_0^{T_f} \left[-\sum_{i=1}^{2} r_i(t) U_i\left(H_i(f_i(t)) + t\right) \right] dt \\ \text{subject to} \quad & \dot{f}_i(t) = -\frac{r_i(t)}{p_i}, i = 1, 2, \\ & f_i(0) = 1, \text{and } f_i(T_f) = 0, \forall\, i = 1, 2, \\ & \mathbf{r}(t) \in \mathcal{C}. \end{aligned} \tag{41}$$

Here, the time to drain the system, T_f is generally not the same for all nonidling scheduling policies. Therefore, this is not a fixed-terminal time problem, rather, the terminal state is specified.

The Hamiltonian for this problem is given by

$$\mathcal{H}(\mathbf{f}(t), \mathbf{r}(t), \mathbf{q}(t)) = -A_1(t) r_1(t) - A_2(t) r_2(t)$$

where$A_i(t) = U_i(D_i(t)) + \frac{q_i(t)}{p_i}$, and the co-state satisfies $\dot{q}_i(t) = r_i \dot{U}_i(D_i(t)) \dot{H}_i(f_i(t))$. Therefore, the optimal control, $\mathbf{r}^*(t)$, satisfies

$$\mathbf{r}^*(t) = \arg\max_{\mathbf{r} \in \mathcal{C}} (-A_1(t) r_1 - A_2(t) r_2) \tag{42}$$

for each time t. As for (40), this always has a solution that lies in $\delta\mathcal{C}$, and if $\mathcal{C}$ has a strictly convex boundary, then (42) has a unique solution for each time t; i.e., there are no singular intervals.

In the case where $\mathcal{C}$ has a strictly convex boundary, the following proposition gives a necessary condition for the $\dot{U}R$ rule to be optimal.

Proposition 3: If the capacity region $\mathcal{C}$ has a strictly convex boundary, and at time $t = 0$, the solution to (40) is in the interior of $\delta\mathcal{C}$, then a necessary condition for the $\dot{U}R$ rule to be optimal is that there exists a constant K such that the $\dot{U}R$ rule gives

$$\dot{U}_1(D_1(t)) = K \dot{U}_2(D_2(t)) \tag{43}$$

for all $t \in [0, T_f]$.

The proof is given in Appendix IV. At $t = 0$, the solution to (40) depends only on the utilities through the ratio $V(0) = \frac{\dot{U}_1(D_1(0))}{\dot{U}_2(D_2(0))}$. The assumption that the solution to (40) is in the interior of $\delta\mathcal{C}$ and that $\delta\mathcal{C}$ is strictly convex implies that there is only one value of $V(0)$ that will give this solution. This proposition then says that the $\dot{U}R$ rule is optimal if and only if the $\dot{U}R$ scheduler gives $V(t) = K$ for all t. This implies that the $\dot{U}R$ rate allocation is fixed for all time t. We also note that the same proof applies if only a portion of δC is strictly convex, as long as the solution to (40) is in the interior of this region at $t = 0$.

As an example, consider a system with uniform initial delays on $[0, 1]$ for each class, and $U_i(D_i) = w_i U(D_i), i = 1, 2,$ where $U(D)$ is the same for both classes and w_i is a class dependent weight. In this case

$$\frac{\dot{U}_1(D_1(0))}{\dot{U}_2(D_2(0))} = \frac{w_1 \dot{U}(1)}{w_2 \dot{U}(1)} = \frac{w_1}{w_2},$$

so that at time $t = 0$, (40) corresponds to maximizing the weighted sum rate $(w_1 r_1 + w_2 r_2)$ for the two classes. If the maximum weighted sum rate is achieved at an interior point of δC, then according to Proposition 3, for the $\dot{U}R$ rule to be optimal, it must give $D_1(t)$ and $D_2(t)$ that satisfy $\dot{U}(D_1(t)) = \frac{w_1}{w_2} \dot{U}(D_2(t))$ for all t. Since the utilities are the same, this implies that $D_1(t) = D_2(t)$ for all t, and so $\dot{f}_1(t) = \dot{f}_2(t)$, or equivalently

$$\frac{r_1}{p_1} = \frac{r_2}{p_2} \tag{44}$$

where r_1 and r_2 are the rates that maximize the weighted sum rate for the two users. In other words, the line $r_1 = \frac{p_1}{p_2} r_2$ must intersect δC at the point that maximizes the weighted sum rate. For a given capacity region and utility weights, this implies that there is only one particular ratio of p_1 and p_2 for which the $\dot{U}R$ rule might be optimal, and this ratio must be "matched" to the utility weights.

Proposition 3 provides a necessary condition for the $\dot{U}R$ rule to be optimal. We have not shown sufficiency of these conditions in general, but we can show this in the following special cases.[17]

Proposition 4: Assume both classes have uniform initial delays on $[0, 1]$ and the same utility function. If the necessary conditions in Proposition 3 are satisfied, then the $\dot{U}R$ rule is optimal in the following cases:

1) The rates selected by the $\dot{U}R$ scheduler satisfy $\frac{r_1}{p_1} = \frac{r_2}{p_2} = 1$.
2) The utilities are affine, i.e., $U(D_i) = a - bD_i$ for some constants a and $b > 0$.

The proof is given in Appendix V. This can be generalized to the case where the initial delays are uniform on any interval $[D^{\min}, D^{\max}]$; the first condition then becomes $\frac{r_1}{p_1} = \frac{r_2}{p_2} = D^{\max} - D^{\min}$.

VII. Conclusion

We have presented an analysis of a simple scheduling rule for a downlink wireless data service, which takes into account the utility derived from each scheduled packet. To maximize the first-order change in utility, the scheduler chooses the packet with the largest product of marginal utility and achievable rate. By assigning different utility functions across users, the scheduler can account for both relative preferences and channel conditions across users.

[17] The difficulty is that the problem is not jointly convex in the control and state variables, which precludes appealing to standard sufficiency results.

We studied the performance of this scheduler for a fluid draining model where the utility is a function of delay. Assigned to each packet are an initial delay and rate, which are chosen independently from the corresponding distributions. In this setting we are able to derive a differential equation, which describes how scheduling resources, or the total service time, is split among the remaining packets as time progresses. The performance with a continuous rate distribution across packets is evaluated by extending the corresponding analysis with a discrete rate distribution. Performance measures such as average utility and delay can be explicitly computed, and a comparison with simulation results shows that the limiting analysis accurately predicts the performance of finite-size systems of interest.

We next looked at the optimal scheduling policy for a system with two classes of users. We formulated this as an optimal control problem in which the objective is to maximize the total utility per packet. Using Pontryagin's minimum principle, we showed that for a system with a TDM capacity region both the optimal and the $\dot{U}R$ scheduling policy must be exactly the same whenever the service time is split between the two classes. For a general utility function, the way in which the optimal scheduler alternates service between the two classes may differ from the $\dot{U}R$ rule. However, we specified conditions on the utility functions, which guarantee that this order will be the same, so that the $\dot{U}R$ rule is optimal. These conditions apply for many utility functions of interest. We also considered the optimal scheduling policy for a non-TDM capacity region with a strictly convex boundary. In that case, we showed that much stronger conditions are needed for the $\dot{U}R$ rule to be optimal. We provided necessary conditions for this to be true and discussed some simple cases where these conditions are also sufficient.

In this work, we have not considered dynamically changing channels and retransmissions, which arise in mobile wireless data systems. The $\dot{U}R$ rule can, in principle, be modified to take these additional features into account. Associated modeling and performance issues are topics for further study.

Appendix I

Proof of Lemma 3: Assume that for a given interval $I = [a, b)$ as in the lemma, $\Delta(t)$ is nonincreasing on I and there exists $t_1 \in I$ such that $\dot{\Delta}(t_1) = 0$. According to (37), the optimal control is $\alpha_1(t_1) = 1$ and $\alpha_2(t_1) = 0$. We will show that this choice of $\alpha_1(t_1)$ implies that there exists a $t_2 \in I$ such that $\dot{\Delta}(t_2) > 0$, creating a contradiction.

If $\alpha_1(t_1)$ fraction of resources is devoted to serving class 1 packets, we have

$$\begin{aligned}\ddot{\Delta}(t_1) &= \dot{M}_2(t_1) - \dot{M}_1(t_1)\\ &= \ddot{U}_1(D_1(t_1))R_1\\ &\quad \times \left[-\dot{H}_1(f_1(t))\frac{\alpha_1(t_1)}{p_1}R_1 + 1\right]\\ &\quad - \ddot{U}_2(D_2(t_1))R_2\\ &\quad \times \left[-\dot{H}_2(f_2(t))\frac{\alpha_2(t_1)}{p_2}R_2 + 1\right].\end{aligned}$$

Taking the derivative with respect to $\alpha_1(t_1)$, and recalling that $\alpha_2(t_1) = 1 - \alpha_1(t_1)$, we have

$$\begin{aligned}\frac{d\ddot{\Delta}(t_1)}{d\alpha_1} &= -\ddot{U}_1(D_1(t_1))\dot{H}_1(f_1(t_1))\frac{R_1^2}{p_1}\\ &\qquad - \ddot{U}_2(D_2(t_1))\dot{H}_2(f_2(t_1))\frac{R_2^2}{p_2}\\ &> 0 \end{aligned} \tag{45}$$

since $U_i(\cdot)$ is concave and $H_i(\cdot)$ is increasing. Let $\alpha_1^*(t_1)$ be the solution to $\ddot{\Delta}(t_1) = 0$, which corresponds to the split given by Theorem 1. Since $U(\cdot)$ is regular, for $|Q(t)| = 2$ (i.e., both classes are being served) we must have $\alpha_1^*(t_1) < 1$. Therefore, from (45), $\ddot{\Delta}(t_1) > 0$, and since $\Delta(t)$ and $\dot{\Delta}(t)$ are both continuous, for a small enough δt, we must have $\dot{\Delta}(t_1 + \delta t) > 0$ and $t_2 = t_1 + \delta t < b$. □

Appendix II

Proof of Lemma 4: Let $t^* = \inf\{t : \Delta(t) \leq 0\}$. From Lemma 3, $\Delta(t)$ must be strictly decreasing for $t \in [0, t^*)$. Since $\dot{\Delta}(t) = M_2(t) - M_1(t)$, we have $M_1(t) > M_2(t)$. Hence, both the optimal scheduler and the $\dot{U}R$ rule schedule class 1 packets for $t \in [0, t^*)$.

At $t^*, \Delta(t^*) = 0$ (otherwise class 1 packets would never be scheduled), and either $\dot{\Delta}(t^*) < 0$ or $\dot{\Delta}(t^*) = 0$. For the first case (see Fig. 8(c)), $\Delta(t) < 0$ for all $t \in (t^*, T_f]$. Therefore, the optimal scheduler serves only class 2 packets for $t \in (t^*, T_f]$. This implies that class 1 packets must be drained at time t^*. The $\dot{U}R$ rule is equivalent to the optimal scheduler since both switch to serve class 2 packets at the same time $t^* = t_2^{\text{in}} = p_1/R_1$.

For the second case, if $\dot{\Delta}(t^*) = 0$ only at the isolated time t^* (see Fig. 8(b)), then $\Delta(t) < 0$ for all $t \in (t^*, T_f]$, and the preceding argument again shows that the $\dot{U}R$ rule is optimal. Now suppose that a singular interval exists where $\dot{\Delta}(t) = 0$ for all $t \in [t^*, s]$ as in Fig. 8(a). From Lemma 3, $t^* = \inf\{t : \dot{\Delta}(t) = 0\}$, and therefore $t^* = t_2^{\text{in}}$. Lemma 2 then states that the $\dot{U}R$ scheduler is optimal for $t \in [t^*, s)$, where $\alpha_1(t)$ is chosen according to (16). For $t > s, \Delta(t) < 0$, so that all class 1 packets must be served at time s, i.e., $s = t_1^{\text{out}}$, and for $t \in [s, T_f)$, both the optimal scheduler and the $\dot{U}R$ rule schedule class 2 packets only. □

Appendix III

Proof of Theorem 2: First we show that conditions i) and ii) jointly imply that the utility function is regular. Given a time t_0 such that $R_1\dot{U}_1(D_1(t_0)) = R_2\dot{U}_2(D_2(t_0))$, let $D_1^1(t_0 + s) = H_1\left(f_1(t_0) - \frac{R_1}{p_1}s\right) + t_0 + s$, for $s \geq 0$, and let $D_2^1(t_0 + s) = D_2(t_0) + s$. Similarly, let $D_1^2(t_0 + s) = D_1(t_0) + s$, and $D_2^2(t_0 + s) = H_2\left(f_2(t_0) - \frac{R_2}{p_2}s\right) + t_0 + s$. Next, define

$$\Theta_1(t_0 + s) = R_1\dot{U}_1(D_1^1(t_0 + s)) - R_2\dot{U}_2(D_2^1(t_0 + s))$$

and

$$\Theta_2(t_0 + s) = R_1\dot{U}_1(D_1^2(t_0 + s)) - R_2\dot{U}_2(D_2^2(t_0 + s)).$$

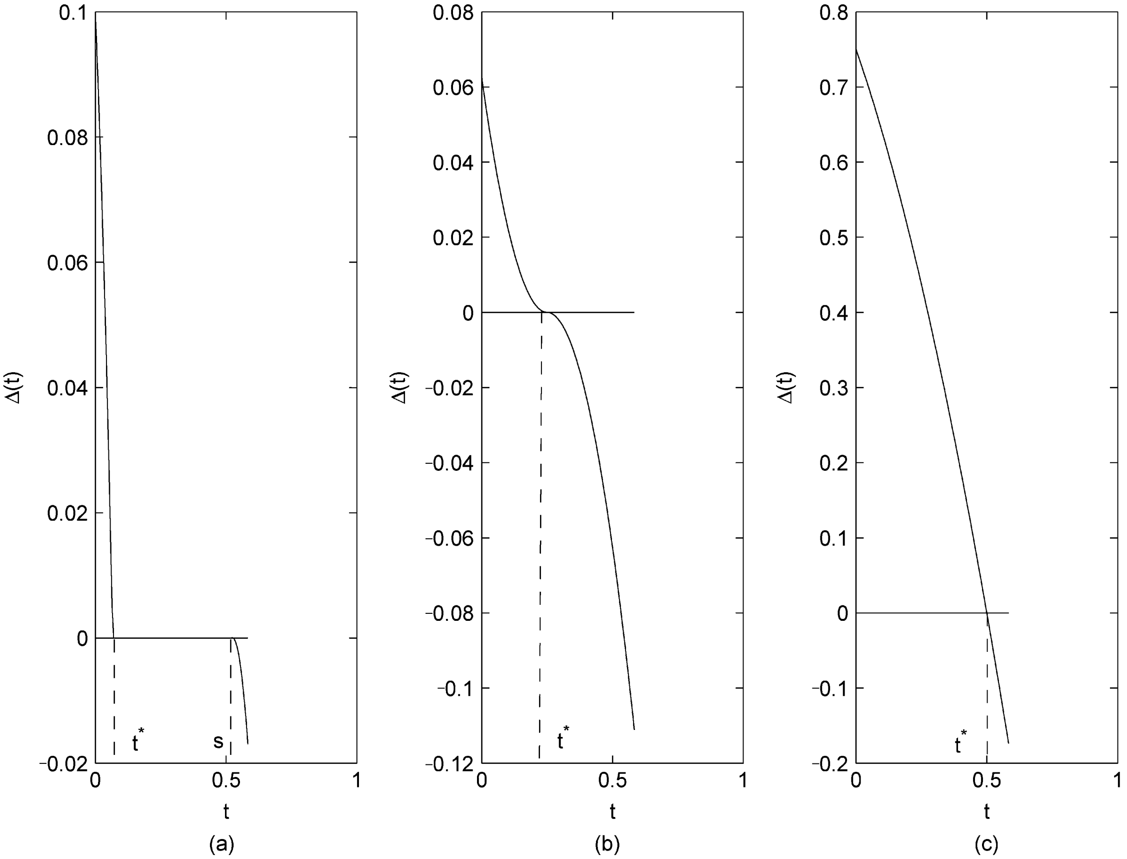


Fig. 8. Possible behaviors of $\Delta(t)$: (a) shows $\Delta(t) = 0$ and $\dot{\Delta}(t) = 0$, for $t \in (t^*, s)$ (a singular interval); (b) shows $\Delta(t^*) = 0$ and $\dot{\Delta}(t^*) = 0$ (no singular interval); and (c) shows $\Delta(t^*) = 0$ and $\dot{\Delta}(t^*) < 0$ (no singular interval).

Note that $\Theta_1(t_0) \equiv \Theta_2(t_0) = 0$. Condition i) in the Theorem states that $\Theta_1(t_0 + s) > 0$ for all $s > 0$. Therefore,

$$\begin{aligned} &\frac{d\Theta_1(t_0+s)}{ds}|_{t=s} \\ &\quad = R_1\ddot{U}_1(D_1(t_0))\left[1 - \dot{H}_1(f_1(t_0))\frac{R_1}{p_1}\right] \\ &\qquad - R_2\ddot{U}_2(D_2(t_0)) \\ &\quad > 0. \end{aligned} \tag{46}$$

Likewise, from condition ii)

$$\begin{aligned} &\frac{d\Theta_2(t_0+s)}{ds}|_{t=s} \\ &\quad = R_1\ddot{U}_1(D_1(t_0)) \\ &\qquad - R_2\ddot{U}_2(D_2(t_0))\left[1 - \dot{H}_2(f_2(t_0))\frac{R_2}{p_2}\right] \\ &\quad > 0. \end{aligned} \tag{47}$$

Let α_1^* be the solution to

$$\begin{aligned} &R_1\ddot{U}_1(D_1(t_0))\left[1 - \dot{H}_1(f_1(t_0))\frac{\alpha_1 R_1}{p_1}\right] \\ &\quad - R_2\ddot{U}_2(D_2(t_0))\left[1 - \dot{H}_2(f_2(t_0))\frac{(1-\alpha_1)R_2}{p_2}\right] \\ &= 0. \end{aligned} \tag{48}$$

Comparing (46), (47), and (48), it follows that (48) must have a unique solution α_1^* that satisfies $0 < \alpha_1^* < 1$. Therefore, the utility function satisfying condition i) and ii) is regular.

Second, we show that the $\dot{U}R$ rule is optimal for any $U(D)$ satisfying the two conditions. First recall that by assumption

$$\dot{\Delta}(0) = \dot{U}_1(D_1(0))R_1 - \dot{U}_2(D_2(0))R_2 \le 0 \tag{49}$$

which implies that the $\dot{U}R$ scheduler begins serving class 1 packets at $t = 0$. Let $t_1 = \inf\{t : \dot{\Delta}(t) > 0\}$, where if $\Delta(t) \le 0$ for all t, then t_1 does not exist. If t_1 does not exist, then $\Delta(t)$ is nonincreasing and the desired result follows from Lemma 4. Therefore, we assume that t_1 exists in the following. From (49) and the continuity of $\dot{\Delta}(t)$, it follows that if t_1 exists then $\dot{\Delta}(t_1) = 0$.

First, we show that if t_1 exists then it must be that $\Delta(t_1) = 0$ (see Fig. 9). Assume that this is not true, so that either $\Delta(t_1) > 0$ or $\Delta(t_1) < 0$.

If $\Delta(t_1) > 0$, then it follows that $\Delta(t) > 0$ for all $t \in [0, t_1]$, and so, the optimal scheduler would serve only class 1 packets for all $t \in [0, t_1]$. However, if $\alpha_1(t_1) = 1$, then from condition i), it follows that both $\dot{\Delta}(t) > 0$ and $\Delta(t) > 0$ for all $t \in (t_1, T_f]$. But this implies that $\Delta(t) > 0$ for all $t \in [0, T_f]$, which cannot be true, since the class 2 packets are never served.

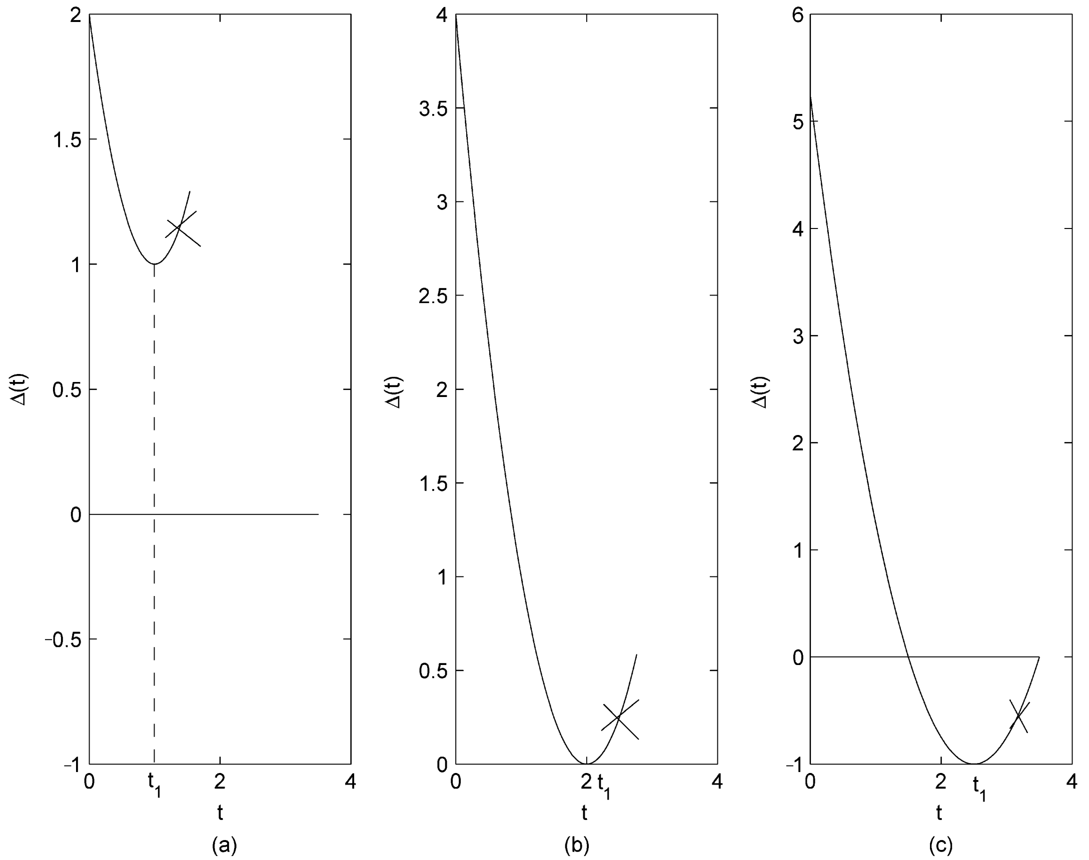


Fig. 9. Example trajectories of $\Delta(t)$ versus t with $\dot{\Delta}(t_1) = 0$: (a) shows $\Delta(t_1) > 0$; (b) shows $\Delta(t_1) = 0$; and (c) shows $\Delta(t_1) < 0$; Both cases (a) and (c) result in contradictions and so can not occur. Case (b) may only occur if $t_1 = t_2^{\text{out}}$.

Next consider the case, where $\Delta(t_1) < 0$. That implies $\alpha_1(t_1) = 0$, and from condition ii), it follows that both $\dot{\Delta}(t) < 0$ and $\Delta(t) < 0$ for all $t \in (t_1, T_f]$. This, however, contradicts the definition of t_1.

Therefore $\Delta(t) \geq 0$ for all $t \leq t_1$. Let $t_2 = \inf\{t : \Delta(t) = 0\}$. For $t \leq t_1$, $\dot{\Delta}(t) \leq 0$, and so $\dot{\Delta}(t) = 0$ for all $t \in [t_2, t_1]$, and from Lemma 3, $\dot{\Delta}(t) < 0$ for all $t < t_2$. It follows that both the $\dot{U}R$ rule and the optimal policy both schedule only class 1 packets during $[0, t_2]$. Also, the interval $[t_2, t_1]$ is a singular interval, and so by Lemma 2 both policies are the same during this interval.[18] We have shown that until time t_1 both the $\dot{U}R$ and optimal policies are the same. For any $t > t_1$, again using condition i), it can be seen that $\Delta(t) > 0$, and so the optimal policy schedules only class 1 packets. Thus, $t_1 = t_2^{\text{out}}$, and so the $\dot{U}R$ policy is optimal. □

Appendix IV

Proof of Proposition 3: Assume that the $\dot{U}R$ is optimal and that at $t = 0$ the solution to (40) is in the interior of $\delta\mathcal{C}$. Under the $\dot{U}R$ rule, the delays and hence the ratio $V(t)$ vary continuously with t. Therefore, there must exist some time $\tilde{t} > 0$ such that the solution to (40) is in the interior of $\delta\mathcal{C}$ for all $t \in [0, \tilde{t})$. If the $\dot{U}R$ rule is optimal, then (40) and (42) must have the same solution (with the appropriate co-state variables). Since the capacity region has a strictly convex boundary, it follows that we must have

$$\frac{\dot{U}_1(D_1(t))}{A_1(t)} = \frac{\dot{U}_2(D_2(t))}{A_2(t)} \tag{50}$$

for all $t \in [0, \tilde{t})$. From their definitions, it can be seen that for $i = 1, 2$

$$\frac{\dot{U}_i(D_i(t))}{A_i(t)} = \frac{d}{dt} \ln\left(U_i(D_i(t)) + \frac{q_i(t)}{p_i}\right).$$

Therefore, there must exist some constant K' such that for all $t \in [0, \tilde{t})$

$$\ln\left(U(D_1(t)) + \frac{q_1(t)}{p_1}\right) - \ln\left(U(D_2(t)) + \frac{q_2(t)}{p_2}\right) = K'.$$

It follows that

$$U(D_1(t)) + \frac{q_1(t)}{p_1} = e^{K'}\left(U(D_2(t)) + \frac{q_2(t)}{p_2}\right)$$

for all $t \in [0, \tilde{t})$. Differentiating and simplifying gives $\dot{U}(D_1(t)) = e^{K'}\dot{U}(D_2(t))$, or $V(t) = e^{K'} \equiv K$ for all $t \in [0, \tilde{t})$. To summarize, we have shown that if the $\dot{U}R$ rule is optimal, then $V(t) = K$ for some $K \geq 0$ for all $t \in [0, \tilde{t})$.

[18] Once again one of these intervals may have measure zero.

Again appealing to continuity, it follows that $V(\tilde{t}) = V(0)$, which corresponds to an interior point of δC, and therefore $\tilde{t} = T_f$. □

Appendix V

Proof of Proposition 4: Assume that we have a system where both classes have uniform initial delays and the same utility function, and assume that the rate allocation under the $\dot{U}R$ rule satisfies the necessary conditions in Prop. 3. Let (r_1^*, r_2^*) be the resulting fixed rate allocation under the $\dot{U}R$ rule. From the discussion following Prop. 3, it follows that these rates maximize the sum capacity and must satisfy (44).

Consider the following optimal control problem:

$$\begin{aligned} \min_{r_1(t), r_2(t)} \quad & \int_0^{T_f} \sum_{i=1}^{2} r_i^* f_i(t) dt \\ \text{subject to :} \quad & \dot{f}_i(t) = -\frac{r_i(t)}{p_i}, \quad i = 1, 2 \\ & f_i(0) = 1, \quad \text{and } f_i(T_f) = 0, \quad \forall\, i = 1, 2 \\ & \mathbf{r}(t) \in \mathcal{C}. \end{aligned} \tag{51}$$

Notice that this is the same as (41), except for a different objective function. The solution to this problem is characterized by the following lemma.[19]

Lemma 5: The solution to (51) is given by $r_i(t) = r_i^*$ for all t.

Proof: The Hamiltonian for (51) is given by

$$\mathcal{H}(\mathbf{f}(t), \mathbf{r}(t), \mathbf{q}(t)) = \sum_{i=1}^{2} r_i^* f_i(t) - \frac{q_i(t)}{p_i} r_i(t)$$

and the co-state equations are

$$\dot{q}_i(t) = -r_i^*,$$

for $i = 1, 2$. Therefore an optimal rate allocation for (51) must satisfy

$$\max_{\mathbf{r}(t) \in \mathcal{C}} \sum_i \frac{q_i(0) - r_i^* t}{p_i} r_i(t) \tag{52}$$

for all t. Consider setting $q_i^*(0) = r_i^*$. Recalling that r_i^* satisfies (44) and maximizes the sum-rate, it follows that the corresponding solution to (52) is $\mathbf{r}(t) = (r_1^*, r_2^*)$ for all time t. So, this choice of the co-state and the corresponding control satisfy the necessary conditions for optimality. Furthermore, in (51) both the objective and the dynamics are jointly convex in $(\mathbf{f}(t), \mathbf{r}(t))$, which implies that the necessary conditions are also sufficient [34]. □

Continuing with the proof of Prop. 4, let $f_i^*(t) = 1 - \frac{r_i^*}{p_i} t$ for $i = 1, 2$. Consider any other feasible rate allocation $\mathbf{r}(t) = (r_1(t), r_2(t))$, and let $f_i(t)$ be the corresponding fraction of remaining packets under this policy. Let $h_i(t) = f_i(t) - f_i^*(t)$, so that $r_i(t) = r_i^* - p_i \dot{h}_i(t)$. Since $\mathbf{r}(t)$ is feasible, it must be that $h_i(0) = h_i(T_f) = 0$ for $i = 1, 2$. The total utility under this rate allocation can be bounded as follows:

$$\begin{aligned} &\int_0^{T_f} \left[-\sum_{i=1}^{2} r_i(t) U(f_i(t) + t) \right] dt \\ &= \int_0^{T_f} -\sum_{i=1}^{2} (r_i^* - p_i \dot{h}_i(t)) U(f_i^*(t) + h_i(t) + t) dt \\ &\geq \int_0^{T_f} -\sum_{i=1}^{2} (r_i^* - p_i \dot{h}_i(t)) \\ &\quad \times \left[U(f_i^*(t) + t) + \dot{U}(f_i^*(t) + t) h_i(t) \right] dt \\ &= -\int_0^{T_f} \sum_{i=1}^{2} r_i^* U(f_i^*(t) + t) dt \\ &\quad + \int_0^{T_f} \sum_{i=1}^{2} p_i \dot{h}_i(t) U(f_i^*(t) + t) dt \\ &\quad - \int_0^{T_f} \sum_{i=1}^{2} r_i^* \dot{U}(f_i^*(t) + t) h_i(t) dt \\ &\quad + \int_0^{T_f} \sum_{i=1}^{2} p_i \dot{h}_i(t) \dot{U}(f_i^*(t) + t) h_i(t) dt \end{aligned} \tag{53}$$

where we have used the fact that $U(D)$ is concave. In (53) we have bounded the total (negative) utility under the rate allocation $\mathbf{r}(t)$ by four terms, the first term being the value obtained from the $\dot{U}R$ policy. To complete the proof, we will show that the remaining three terms are all greater than or equal to zero. We consider the two cases in the Proposition separately.

Case 1: $\frac{r_1^*}{p_1} = \frac{r_2^*}{p_2} = 1$

In this case, $\dot{f}_i^*(t) = -1$ and so $f_i^*(t) + t = 1$ for all t and for each class i. Therefore the terms $U(f_i^*(t) + t)$ and $\dot{U}(f_i^*(t) + t)$ in (53) are constants for all t. For the second term in (53) we then have

$$\begin{aligned} &\int_0^{T_f} \sum_{i=1}^{2} p_i \dot{h}_i(t) U(f_i^*(t) + t) dt \\ &= U(1) \sum_{i=1}^{2} \left(p_i \int_0^{T_f} \dot{h}_i(t)\, dt \right) \\ &= U(1) \sum_{i=1}^{2} p_i \left(h_i(T_f) - h_i(0) \right) \\ &= 0. \end{aligned}$$

Likewise, for the fourth term in (53) we have

$$\begin{aligned} &\int_0^{T_f} \sum_{i=1}^{2} p_i \dot{h}_i(t) \dot{U}(f_i^*(t) + t) h_i(t) dt \\ &= \dot{U}(1) \sum_{i=1}^{2} p_i \left(\int_0^{T_f} h_i(t) \dot{h}_i(t) dt \right) \\ &= \dot{U}(1) \sum_{i=1}^{2} p_i (h_i(T_f))^2 - (h_i(0))^2 \\ &= 0. \end{aligned}$$

[19] Note this lemma only applies under the current assumptions of uniform initial delays, the same utility for both classes, and that the rate allocation (r_1^*, r_2^*) under the $\dot{U}R$ rule satisfies the necessary conditions.

Finally, for the third term in (53) Lemma 5 states that the fixed rate allocation (r_1^*, r_2^*) minimizes (51) over all feasible rate allocations. From this it follows that

$$-\int_0^{T_f} \sum_{i=1}^{2} r_i^* \dot{U}(f_i^*(t)+t) h_i(t) dt = (-\dot{U}(1)) \int_0^{T_f} \sum_{i=1}^{2} r_i^* h_i(t) dt \geq 0$$

and so the third term in (53) must be nonnegative, which completes the proof for the first case.

Case 2: $U(D) = a - bD$:

In this case $\dot{U}(D_i) = -b$ is a constant; therefore we can use the same argument as in the first case to bound the fourth term in (53). For the second term in (53), we have

$$\begin{aligned}
&\int_0^{T_f} \sum_{i=1}^{2} p_i \dot{h}_i(t) U(f_i^*(t)+t) dt \\
&= \int_0^{T_f} \sum_{i=1}^{2} p_i \dot{h}_i(t) \left[a - b \left(1 - \frac{r_i^*}{p_i} t + t \right) \right] dt \\
&= \sum_{i=1}^{2} p_i \left[(a-b) \int_0^{T_f} \dot{h}_i(t) dt \right. \\
&\qquad \left. - b \left(1 - \frac{r_i^*}{p_i} \right) \int_0^{T_f} \dot{h}_i(t) t dt \right] \\
&= \sum_{i=1}^{2} \left(-b \left(p_i - r_i^* \right) \int_0^{T_f} \dot{h}_i(t) t dt \right) && (54) \\
&= \sum_{i=1}^{2} \left(-b \left(r_i^* - p_i \right) \int_0^{T_f} h_i(t) dt \right) && (55)
\end{aligned}$$

where (54) follows by the same argument as in case 1, and (55) follows from integrating by parts and using the fact that $h_i(0) = h_i(T_f) = 0$. Combining this with the third term in (53) yields the term $b \int_0^{T_f} \sum_{i=1}^{2} p_i h_i(t)\, dt \equiv y$. Since $\frac{r_1^*}{p_1} = \frac{r_2^*}{p_2}$, it follows that the fixed rate allocation (r_1^*, r_2^*) also minimizes $\int_0^{T_f} \sum_{i=1}^{2} p_i f_i(t) dt$ over all feasible rate allocations, and so $y \geq 0$. The desired result for case 2 then follows. □

REFERENCES

[1] L. Tassiulas and A. Ephremides, "Dynamic server allocation to parallel queue with randomly varying connectivity," *IEEE Trans. Inf. Theory*, vol. 39, pp. 466–478, Mar. 1993.

[2] P. Bhagwat, P. Bhattacharya, A. Krishna, and S. K. Tripathi, "Enhancing throughput over wireless LAN's using channel state dependent packet scheduling," in *Proc. INFOCOM*, San Franciso, CA, Mar. 1996, pp. 1133–1140.

[3] V. Bharghavan, S. Lu, and T. Nandagopal, "Fair queuing in wireless networks: Issues and approaches," *IEEE Pers. Commun.*, vol. 6, pp. 44–53, Feb. 1999.

[4] M. Andrews, K. Kumaran, K. Ramanan, A. L. Stolyar, R. Vijayakumar, and P. Whiting, "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, vol. 39, no. 2, pp. 150–154, 2001.

[5] M. Andrews, K. Kumaran, K. Ramanan, A. L. Stolyar, R. Vijayakumar, and P. Whiting, "Scheduling in a queueing system with asynchronously varying service rate," *Probability in the Engineering and Informational Sciences*, vol. 18, pp. 191–217, 2004.

[6] R. Agrawal, A. Bedekar, R. La, and V. Subramanian, "A class and channel-condition based weighted proportionally fair scheduler," in *Proc. ITC 2001*, Salvador, Brazil, Sep. 2001.

[7] R. Agrawal and V. Subramanian, "Optimality of certain channel aware scheduling policies," in *Proc. 2002 Allerton Conf. Commun., Contr. Comput.*, Oct. 2002.

[8] R. Leelahakriengkrai and R. Agrawal, "Scheduling in multimedia CDMA wireless networks," *IEEE Trans. Veh. Technol.*, 2002.

[9] S. Shakkottai and A. L. Stolyar, "Scheduling algorithms for a mixture of real-time and nonreal-time data in HDR," in *Proc. 17th Int. Teletraffic Congr.*, Salvador da, Bahia, Brazil, Sept. 24–28, 2001, pp. 793–804.

[10] S. Shakkottai and A. L. Stolyar, "Scheduling for multiple flows sharing a time-varying channel: The exponential rule," *Anal. Meth. Appl. Prob.*, ser. 2, vol. 207, American Mathematical Society Translations, pp. 185–202, 2002.

[11] S. Shakkottai and R. Srikant, "Scheduling real-time traffic with deadlines over a wireless channel," *ACM/Baltzer Wireless Netw. J.*, vol. 8, no. 1, pp. 13–26, Jan. 2002.

[12] X. Liu, E. K. P. Chong, and N. Shroff, "Opportunistic transmission scheduling with resource sharing constraints in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 10, Oct. 2001.

[13] Y. Liu and E. Knightly, "Opportunistic fair scheduling over multiple wireless channels," in *Proc. IEEE INFOCOM*, 2003.

[14] H. Kushner and P. Whiting, "Asymptotic properties of proportional-fair sharing algorithms," in *Proc. 2002 Allerton Conf. Commun., Contr. Comput.*, Oct. 2002.

[15] K. Lee and M. El Zarki, "Scheduling real time traffic in IP-based cellular network," in *Proc. IEEE Int. Symp. Personal, Indoor and Mobile Radio Communications (PIMRC)*, London, U.K., Sep. 2000, pp. 1202–1206.

[16] P. Liu, R. Berry, and M. Honig, "Delay-sensitive packet scheduling in wireless networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, New Orleans, LA, Mar. 2003.

[17] R. Berry, P. Liu, and M. Honig, "Design and analysis of downlink utility-based schedulers," in *Proc. 40th Ann. Allerton Conf. Commun., Contr. Comput.*, Monticello, IL, Oct. 2002.

[18] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi, "CDMA/HDR: A bandwidth-efficient high-speed wireless data service for nomadic users," *IEEE Commun. Mag.*, pp. 70–77, July 2000.

[19] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency—High data rate personal communication wireless system," in *Proc. VTC '2000*, Spring 2000.

[20] *TIA/EIA IS-856 CDMA 2000: High Rate Packet Data Air Interface Specification*, , Nov. 2000.

[21] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beam-forming using dumb antennas," *IEEE Trans. Inf. Theory*, vol. 48, June 2002.

[22] J. A. Van Mieghem, "Dynamic scheduling with convex delay costs: The generalized $c\mu$ Rule," *Ann. Appl. Prob.*, vol. 5, no. 3, 1995.

[23] A. Mandelbaum and A. L. Stoylar, "$GC\mu$ scheduling of flexible servers: Asymptotic optimality in heavy traffic," in *Proc. 2002 Allerton Conf. Commun., Contr. Comput.*, Oct. 2002.

[24] R. Haji and G. F. Newell, "Optimal strategies for priority queues with nonlinear costs of delay," *SIAM J. Appl. Math.*, vol. 20, no. 2, pp. 224–240, Mar. 1971.

[25] A. L. Stolyar, "MaxWeight scheduling in a generalized switch: State space collapse and equivalent workload minimization in heavy traffic," *Ann. Appl. Prob.*, vol. 14, no. 1, pp. 1–53, 2004.

[26] G. Weiss, "On optimal draining of fluid re-entrant lines," in *Stochastic Networks: IMA Volumes in Mathematics andits Applications*, F. P. Kelly and R. Williams, Eds. New York: Springer-Verlag, 1995, pp. 93–105.

[27] D. R. Cox and W. Smith, *Queues*. New York: Wiley, 1961.

[28] W. E. Smith, "Various optimizers for single-stage production," *Nav. Res. Logist. Quart.*, vol. 3, pp. 59–66, 1956.

[29] R. Righter, "Scheduling," in *Stochastic Orders*, M. Shaked and J. Shanthikumar, Eds. New York: Academic, 1994.

[30] R. Dudley, *Real Analysis and Probability*. New York: Chapman and Hall, 1989.

[31] F. Riesz and B. Nagy, *Functional Analysis*. New York: Ungar, 1955.

[32] M. Athans and P. Falb, *Optimal Control, An Introduction to the Theory and its Applications*. New York: Mc Graw-Hill, 1966.

[33] D. E. Kirk, *Optimal Control Theory an Introduction*. Englewood Cliffs, NJ: Prentice-Hall, 1970.

[34] M. Kamien and N. Schwartz, *Dynamic Optimization: The Calculus of Variations and Optimal Control in Economics and Management*. New York: Elsevier Science, 1981.